Journal of Development Economics 43 (1994) 5–38. North-Holland

New goods, old theory, and the welfare costs of trade restrictions

Paul Romer*

University of California, Berkeley CA, USA, NBER, and CIAR

Received May 1992, final version received April 1993

The typical economic model implicitly assumes that the set of goods in an economy never changes. As a result, the predicted efficiency loss from a tariff is small, on the order of the square of the tariff rate. If we loosen this assumption and assume that international trade can bring new goods into an economy, the fraction of national income lost when a tariff is imposed can be much larger, as much as two times the tariff rate. Much of this paper is devoted to explaining why this seemingly small change in the assumptions of a model can have such important positive and normative implications. The paper also asks why the implications of new goods have not more extensively been explored, especially given that the basic economic issues were identified 150 years ago. The mathematical difficulty of modeling new goods has no doubt been part of the problem. An equally, if not more important stumbling block has been the deep philosophical resistance that humans feel toward the unavoidable logical consequence of assuming that genuinely new things can happen and could have happened at every date in the past. We are forced to admit that the world as we know it is the result of a long string of chance outcomes.

Key words: Trade restrictions; New goods

JEL classification: F12; F43; O10

1. Introduction

Most economic theory starts from an implicit assumption that policy interventions do not affect the set of goods available in the economy. Recent theoretical work shows that this assumption severely restricts our analysis of growth in advanced economies. We can not ask how policy affects the aggregate rate of invention and innovation if we assume from the outset that no new goods can be introduced. But the assumption that the set of goods is

Correspondence to: Paul Romer, NBER, 204 Junipero Serra Blvd., Stanford, CA 94305, USA.

*An early version of this paper was presented at the NBER Inter-American Seminar in Buenos Aires, Argentina in May 1992. Comments by Sebastian Edward, Felipe Larrain, and conference participants are gratefully acknowledged. Comments by the members of the Program on Economic Growth of the Canadian Institute for Advanced Research were also helpful in the early stages of this work, and detailed comments from James Fox on the final draft were very useful. Research support was provided by the Institute for Policy Reform and NSF grant no. SES 90-23469.

SSDI 0304-3878(93)E0035-V

fixed is just as restrictive for the analysis of growth in developing countries. In particular, it keeps us from studying how trade restrictions prevent new types of goods and new types of productive activities from being introduced from abroad. The ultimate claim in this paper is that a theoretical perspective implicitly based on the idea that no new goods can ever be introduced leads to a substantial underestimate of the welfare costs of trade restrictions.

The discussion leading up to this final claim touches on larger issues surrounding the concept of 'newness' and the lasting contribution that 'new' growth theory can make to economic analysis. The analysis of trade presented below is motivated by the neo-Schumpeterian models of growth that have been developed in the last few years. These models explicitly allow for the introduction into an economy of new or improved types of goods. Early contributions to this branch of growth theory include Aghion and Howitt (1990), Grossman and Helpman (1991, 1992), Romer (1987, 1990) and Segerstrom et al. (1990). This branch of endogenous growth theory differs from the models in Lucas (1988) and Romer (1986), which emphasize external increasing returns, and from models in Jones and Manuelli (1990) and Rebelo (1991), which invoke perfect competition and a broad concept of capital that can be accumulated forever without driving its marginal product to zero. Both the external effects and perfect competition models of endogenous growth maintain the assumption that new goods do not matter at the aggregate level.

The premise in the neo-Schumpeterian models is that every economy faces virtually unlimited possibilities for the introduction of new goods. Advanced nations can discover new goods. Developing countries can import them. The term 'good' is used here in the broadest possible sense. A new good could take the form of an entirely new type of physical good – the digital computer in the 1950s. A new good could also be a quality improvement over an existing physical good – this year's generation of more powerful personal computers. With no fundamental change in the underlying economic analysis, new goods can be modeled either as consumption goods, as in Grossman and Helpman (1992), or as inputs in production, as in Romer (1990).

We tend to think of goods as tangible objects, but clearly they need not be. When a software engineer writes a computer program, she produces a new good. Less obviously, when a process engineer finds a better way to manufacture a product or a manager finds a better way to monitor inventories and distribute goods, they also produce an insight or discovery that is a new good in all of the relevant senses of the word. These discoveries have economic value and are costly to produce. The language used to describe the neo-Schumpeterian models emphasizes tangible goods, but the logic applies equally well to intangibles.

If there are almost limitness numbers of conceivable goods that can be introduced into any economy, there must also be some fixed cost associated with the introduction of each new good. Otherwise, every valuable good would already be in use everywhere. The modeling innovation in the neo-Schumpeterian models of growth is that they take explicit account of the fixed costs that limit the set of goods and show that these fixed costs matter in a dynamic analysis conducted at the level of the economy as a whole. This contrasts with the standard approach in general equilibrium analysis, in which fixed costs are assumed to be of negligible importance in markets of realistic size.

There is, of couse, an extensive literature in industrial organization that takes fixed costs seriously, including a specialized literature on patent races that emphasizes the costs associated with the introduction of new goods. Macroeconomists and general equilibrium theorists who work at the level of the economy as a whole seem not to be impressed with the importance of the lessons from these microeconomic analyses. The neo-Schumpeterian growth models stress just one of the assumptions from the microeconomic literature, that there are fixed costs. These new growth models do not capture the complicated strategic interactions that emerge when there are only a small number of firms in a market. The models nevertheless show that the presence of fixed costs is sufficient to overturn important parts of the conventional wisdom concerning positive and normative analysis at the aggregate level.

In particular, the neo-Schumpeterian models emphasize a point that should already have been clear from the preceding work in new trade theory. [See, for example, Helpman and Krugman (1985).] Trade policy can have large positive and normative effects because it can influence not just quantities of existing goods as traditional trade theory suggests, but also the number of different types of goods that are available in an economy. [For example, Feenstra (1992) gives a calculation of welfare losses from trade restrictions that takes account of these effects.] This important implication of new trade theory has sometimes been overshadowed by an emphasis on the ex post monopoly rents that also arise in models with fixed costs.

Because the traditional argument for free trade relies on perfect competition and because perfect competition cannot be sustained when there are important fixed costs, fixed costs have been used to justify all manner of government intervention, including trade restrictions. Nevertheless, the ultimate claim of the paper is that taking account of fixed costs actually strengthens the arguments in favor of free trade.

Formally, it is true that the equilibria in the neo-Schumpeterian models are not Pareto optimal, and by definition, this raises the theoretical possibility that some form of collective action could improve on decentralized outcomes. This does not, however, imply that there is a feasible policy that a real government could implement that would lead to a Pareto improvement. Even if one believes that the goverment can improve on market outcomes, this does not mean that trade policy is the right tool for trying to do so. And even if one grants the need for intervention and even if trade policy is the only tool the government can use, it does not follow that trade restrictions move the economy in the right direction.

The argument developed here, that trade restrictions can be very harmful, is an application of a familiar result from second-best analysis. If the economy starts at a position that is not first-best Pareto optimal, an intervention that moves the economy in the right direction will have first-order effects that increase welfare. But interventions that move the economy in the wrong direction will have first-order effects that reduce welfare. In contrast, if the economy starts from an equilibrium that is first-best Pareto optimal, all interventions have effects that are second-order small. As a result, the assertion that an equilibrium is not first-best Pareto optimal does not validate any arbitrary intervention. It just raises the stakes. Trade restrictions are a little bit harmful in the usual model of perfect markets. They are very harmful in the second-best world with fixed costs described below.

The first irony inherent in the analysis is therefore that a model of imperfect markets captures more accurately the true welfare costs of trade restrictions than does a model of perfect markets. A second irony stems from the observation that new growth theory is about newness (or at least the neo-Schumpeterian branch of new growth theory is about newness), but the theory itself is not new. The term 'neo-Schumpeterian' has the advantage that it acknowledges Joseph Schumpeter's emphasis of the fixed costs associated with intentional invention and innovation, but the roots of the analysis of new goods can be traced back to much earlier work. In one of the very first uses of the concept of a demand curve, the French engineer Jules Dupuit (reprinted in 1969) outlined the essential points: New goods are associated with fixed costs, and fixed costs pose serious difficulties for decentralized market allocation schemes.

Dupuit wrote about new goods more than 150 years ago. Economists are inundated with new goods in their daily lives. It is therefore somewhat puzzling that the potential for new goods still plays such a small role in aggregate economic analysis. In its discussion of the deeper implications of newness, this paper outlines two different forces that may have tended to keep newness in the background. The most obvious restraining force is the technical difficulty of constructing economy-wide mathematical models with fixed costs. The importance of mathematical difficulty has been noted before. [See, for example, the introduction in Krugman (1990) or the initial sections of Romer (1991).] New goods, fixed costs, and market power are relatively easy to capture in a partial equilibrium model, but much harder to incorporate in analysis conducted at the level of the economy as a whole.

Yet in the last 15 years, we have made significant progress in overcoming the technical limitations imposed by our mathematical tools. In industrial organization, trade, and growth, we have been able to build on the model of differentiated products introduced by Dixit and Stiglitz (1977) and by now have accumulated quite a bit of experience with aggregate-level analysis in models with fixed costs and market power. If we now have the technical tools to handle economy-wide models with nonconvexities, why is it that the potential for new goods has not yet been widely incorporated into economic analysis?

The arguments presented below suggest that there may be a second difficulty that we face in coming to terms with the enormous potential for new things to happen, one that economic historians [for example, David (1985)] and a few abstract theorists [for example, Arthur (1989)] have emphasized for some time. Once we admit that there is room for newness – that there are vastly more conceivable possibilities than realized outcomes – we must confront the fact that there is no special logic behind the world we inhabit, no particular justification for why things are the way they are. Any number of arbitrarily small perturbations along the way could have made the world as we know it turn out very differently.

These kinds of abstract concerns may seem to be far removed from the problems faced by someone who offers advice about economic policy decisions in a developing country. But as section 7 of this paper will show, the implicit assumptions we adopt because of mathematical familiarity and philosophical predisposition can decisively influence our analysis of a practical question such as how trade restrictions affect an economy.

Section 2 of the paper starts by illustrating some of the strengths and weaknesses inherent in the use of formal mathematics in economic theory. The discussion suggests that when general equilibrium theory removed the distinction between static and dynamic analysis, the benefits from this theoretical unification were accompanied by a loss in the appreciation that some verbal theorists had for the almost infinite possibilities presented to us by the physical world. This section shows how the apparently innocuous convexity assumptions adopted by general equilibrium theorists made it impossible for economists to study the emergence of new goods in any interesting sense. Section 3 shows how our neglect of new goods leads to a corresponding blind spot in the partial equilibrium analysis of surplus triangles. Section 4 then expands on the possible role that our philosophical predispositions play in influencing how we think about the world. It argues that what philosophers have called the principle of plenitude exercises a much stronger influence over how economists approach the world than we might care to admit.

Section 5 recapitulates the economic analysis of the decision to build a new bridge that was first presented by Dupuit, and section 6 puts Dupuit's partial equilibrium insights into the kind of explicit, general equilibrium framework that is necessary if we are to draw conclusions about national economic policy. In some ways, section 6 is intended as a partial vindication of the use of formal mathematics in economic theory. It is probably true that formal methods encouraged the profession to pull back from difficult and challenging questions when they were first introduced. As we developed our mathematical skills, we focused on the simple issues that lent themselves most readily to mathematical analysis. But now, mathematical theory has developed to the point where it can uncover points that Dupuit and the verbal theorists who followed after him missed. Specifically, section 6 shows that a familiar partial equilibrium insight about perfect price discrimination does not carry over to a model of the economy as a whole. In a general equilibrium setting, price discrimination or multi-part pricing can not resolve the decentralization problem associated with the introduction of new goods. There is no simple, Pareto efficient, decentralized solution to the problem of deciding which of the many conceivable new goods should be introduced into an economy.

The larger implication that follows from this chain of arguments should be sobering not only for economists accustomed to the viewpoint that economic behavior can be represented as a Pareto optimal outcome of a simple model, but also for economists who believe that the government stands ready to improve on market outcomes. Despite the demonstration that it is theoretically possible to improve on the no-intervention outcome in economies where new goods are important, it is not clear that any actual government will be able or willing to undertake policies that are welfare improving. What is clear is that many governments intervene in ways that substantially reduce welfare. Section 7 gives an illustrative calculation designed to show why trade restrictions, one of the most common forms of government intervention in the developing world, might be far more costly than the traditional analysis, with its fixed set of goods, has led us to believe.

2. General equilibrium analysis and new goods

Changing to models that allow for new goods requires a subtle but important shift in the unexamined assumptions and habits of thought that we bring to any problem. By their very nature, these habits and assumptions are things that we take for granted, so it takes an effort to bring them to the surface and subject them to analysis. To do this, it helps to go back prior to the introduction of formal equilibrium theory and retrace some of the logical steps that led to the unification between dynamic analysis and static analysis that mainstream economists now take as obvious.

When Alfred Marshall, Joseph Schumpeter, or other pre-general equilibrium theorists described the methods of economic analysis, they routinely distinguished static analysis from dynamic analysis. This distinction was

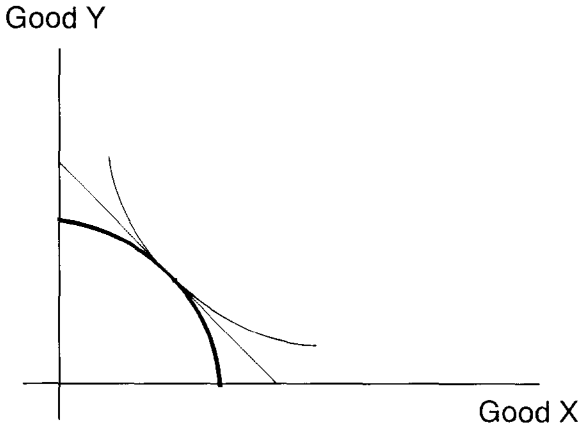

Fig. 1

mirrored in the distinction that economists made between equilibrium states and processes characterized by disequilibrium. This distinction is still used by economists such as Schultz (1975) who were trained prior to the advent of general equilibrium theory and by others who have deliberately moved outside of mainstream mathematical economic theory, such as Nelson and Winter (1982) or the contributors to the volume edited by Dosi et al. (1988).

After general equilibrium theorists gave their description of a general economic equilibrium with dated goods and state-contingent goods, the apparent distinction between statics and dynamics disappeared from equilibrium analysis, as did the distinction between certain and uncertain outcomes. Suddenly it became clear that the methods and modes of analysis that economists could apply to the problem illustrated in fig. 1 are the same regardless of the labels on the axes. The choice illustrated in the figure could be between apples and oranges today. Or it could be between apples today and apples tomorrow. It could even be between oranges tomorrow if there is a frost and oranges tomorrow if there is not.

This unified approach to dated, state-contingent goods ultimately led to a fundamentally new set of mental habits and presumptions. Over time, the gulf between this new point of view and the earlier one became so large that it is now difficult for the economists on the two sides to carry on a sustained intellectual discussion. For example, mainstream mathematical economists and macroeconomists now understand an equilibrium to be whatever happens. The concept no longer carries any presumption of stationarity. Just as there is no reason for the number of apples produced and consumed to be the same as the number of oranges, there is no reason in a dynamic equilibrium for the quantity of goods produced and consumed today to be equal to the comparable quantities at future dates. That is, there is nothing about fig. 1 that requires a tangency point on the 45 degree line. In its most general form, an economic equilibrium could be characterized by slow adjustment, missing markets, nonconvexities, uncertainty, asymmetric information, chaotic dynamics, or the like. Whatever takes place, it is by definition an equilibrium outcome. For people used to looking at the world this way, it simply does not make sense to talk about disequilibrium behavior.

Yet the economists who insist on the importance of disequilibrium phenomena cannot be dismissed as merely being confused. Nor is the disagreement here merely a semantic difference about how the word 'equilibrium' should be defined and used. Economists who view the world from the new general equilibrium point of view are aware of how much insight we have gained from the unifying power of the concepts such as state-contingent goods and dated goods. In this they are surely right.

But economists on the other side of the gulf see that something important was lost in the translation of traditional insights into mathematical terms. These critics are at least partially right, but the difficulty does not arise merely from giving a new meaning to the old notion of an equilibrium. Nor does the difficulty lie with the unification of statics and dynamics per se. Treating apples today and apples tomorrow as being just like apples today and oranges today raises difficult questions about the ability of economic agents to form expectations about future contingencies, but the increased emphasis on learning by advocates of both rational expectations and bounded rationality suggests that disagreements about expectations and rationality may be diminishing.

Instead, the most serious limitation inherent in the general equilibrium approach comes from the convexity assumptions that preclude an analysis of the possibility that many valuable new goods could be, but are not, produced in an economy. Asserting that a new good could be introduced is equivalent to asserting that the economy is currently on the boundary or edge of goods space. Fig. 2 illustrates the only sense in which this can occur if production possibilities sets (and preferences) are convex. According to the case illustrated in this figure, no amount of good S is produced because no amount would be worth the cost in foregone units of good X that it would take to produce any of good S. In fact, good S hardly qualifies as a 'good' at all. If we are not endowed with any amount of good S, we would certaintly not produce any of it, so economists can without loss of generality pretend that good S simply does not exist. If there is a third good Y, which like X, is produced in positive quantities, there is no harm in merely removing the good S from the analysis and proceeding on the assumption that the equilibrium in the economy takes place in the interior of the X–Y goods space, exactly as depicted in fig. 1. With convexity, we can assume that all of

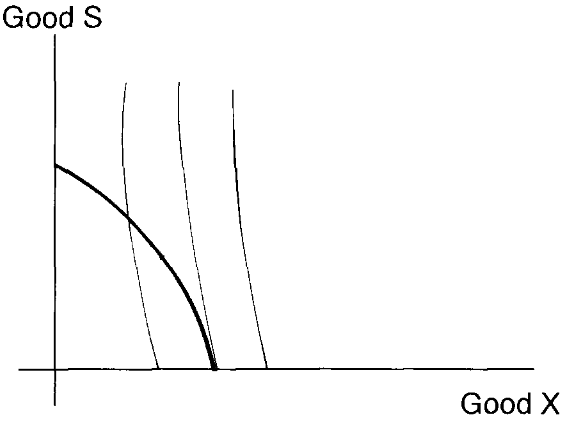

Fig. 2

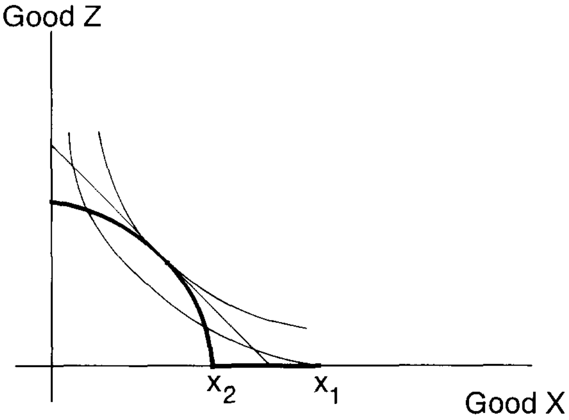

Fig. 3

the relevant goods already exist and assume away all goods that are not produced. We can therefore assume that we are always in the interior of goods space.

Fig. 3 illustrates how a valuable good can exist without being produced in a decentralized equilibrium. The figure requires, however, a nonconvexity that takes the form of a fixed cost. The production set in this figure illustrates the assumption that production of good Z can be made possible at a cost of $x_1 - x_2$ units of good X. Suppose for now that the economy is endowed with X but does not produce any amount of good Z. Suppose also that a positive quantity of a third good Y is also being produced. If economists followed the procedure outlined above – removing from consideration all goods that are not already being produced and that are not part of the endowment bundle – they would remove from consideration an alternative that is relevant for this economy. As fig. 3 is drawn, utility for the representative agent in this economy would be higher if good Z was introduced and a large enough quantity was produced.

The fundamental premise in this paper is that the presence of a potentially valuable good like good Z that has not yet been introduced into the economy is not the exceptional case. Rather, it is the goods that have been introduced that are exceptional. Every real economy is presented with an almost incomprehensible number of new goods that can be introduced. Some of these goods are like good Z in fig. 3. They would increase utility. Many others, perhaps the great majority of all possible new goods, would not be worth introducing. The fixed costs are too high and the benefits too low. Out of the enormous set of possible new goods, a very small number are somehow selected and introduced. In some overall sense, the problem represented by fig. 1 – the relatively simple problem emphasized in most of economic theory of deciding between different quantities of existing goods – is far less common and far less important than the problem illustrated in fig. 3. The economy must decide whether each potential new good is worth the cost it takes to bring it into existence. And the simple diagrammatic analysis offered here vastly understates the complexity of the decision to introduce a good, because the value of any particular new good will depend through complicated chains of complementarity and substitution on the other goods that are present.

Convexity assumptions, which appear to be made purely for technical convenience, therefore have the substantive effect of removing from consideration the most challenging problem confronting an economy. They also limit the ability of a dynamic model to explain basic facts such as that we have personal computers and our grandparents did not. Dynamic general equilibrium models with state contingent goods and convex production set may be useful for some purposes, but the critics are right that there is something fundamental and important about the evolution of an economy that equilibrium models based on convex sets cannot capture.

3. Partial equilibrium analysis and new goods

Many economists would argue that the actual influence of general equilibrium theory is quite limited and that in day to day reasoning, most economists rely primarily on supply and demand curves. The Rorschach test for the attitude that these economists have toward new goods is to present them with fig. 4 and ask them to describe the first thoughts that come to mind. The figure can represent a markup of price over marginal cost that

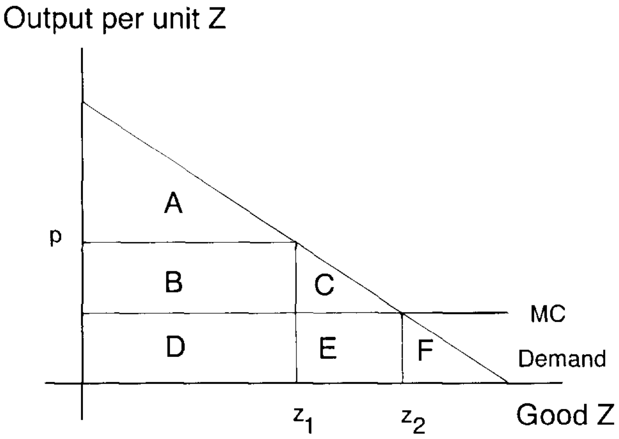

Fig. 4

results from market power. It could also describe a competitive industry with a tax distortion between price and marginal cost. To keep things simple, set aside the niggling disputes about consumer surplus as a welfare measure. Assume that the demand is derived from a production process and prices are measured in units of output from this process. In this case, the demand curve is a marginal productivity schedule. Any areas underneath the curve can rigorously be interpreted as having units of output. To be specific, suppose that the demand curve represents the marginal productivity of some good Z in producing another good X. The marginal cost curve then represents the opportunity cost, in units of the output good X when resources are diverted into the production of good Z.

The interesting question in this version of a psychologist's ink blot test concerns the policy conclusions that an economist attaches to the different labeled areas under the curve. In textbook treatments, and also in the unguarded reaction of many economists, the familiar deadweight triangle labeled C in the figure is the primary focus of attention. Sometimes attention turns as well to the net revenue rectangle B. In a first pass analysis, the redistribution of wealth associated with this rectangle is often ignored. In a second pass, circumstances in which redistribution is a genuine policy concern may be noted. For example, if the demand curve represents demand from the rest of the world and the difference between price and marginal cost is an export tariff, the revenue rectangle represents an increase in national income that comes at the expense of foreigners. (This optimal tariff issue was emphasized in classical trade theory.) Alternatively, the figure could represent the price and quantity decisions of a firm with market power that sells to the rest of the world. (This monopoly rent issue tends to be emphasized in discussions of strategic trade policy that arise from results in new trade theory.) In either case, the revenue rectangle represents a net distribution of income in favor of domestic citizens.

The randomly selected economist taking this version of the ink blot test probably will not cite the triangle A as having any particular welfare significance. This triangle represents a pure surplus of some sort, the recipient of which may not at first be apparent. In any event, it is almost always understood to be of no policy concern. The market creates a surplus, but that is what markets are supposed to do.

But this surplus ought to make economists pause. When a quantity of inputs z_1 is used in the production process which generates the marginal productivity schedule, it produces an amount of output equal to the area $A+B+D$. The market, however, values the units of Z at $z_1p=B+D$. The market rewards the bundle of goods z_1 by paying it only a portion of the output that it produces, and this will be true regardless of the amount of good Z in the bundle.

In competitive markets, prices work at the margin. If good Z already existed, then prices that are equal to marginal cost give the right signals about how much of Z to use in this production process. But these prices do not attach the correct overall value to the associated bundle of goods, and cannot be used as a guide in the decision about whether or not to incur a cost and invent good Z.

This divergence between the social and market valuations of this bundle is a real divergence with potentially important welfare implications. It can be neglected in the standard analysis only because the standard analysis always assumes that the good Z already exists and that only the marginal question matters. If the world is as depicted in fig. 1 – if goods Z and X both exist – then the usual deadweight triangle is the only one that need concern us. All of the action would indeed take place at the margin, in the interior of goods space. But if the world is as depicted in fig. 3, the important action is on the boundary. We must decide whether to incur some fixed cost to make it possible to produce good Z. The total cost of producing z_1 units of good Z will be the fixed cost or introducing good Z plus the marginal cost represented by area D. In deciding whether to introduce the bundle Z, this cost should be compared with the total amount of output that it produces (area $A+B+D$) not with the value attributed to it by the price system (area $B+D$.)

4. The principle of plenitude

The convenient but very powerful assumption that we never have to face the decision about whether to invent a good like good Z in figs. 3 or 4 is a special case of a general presumption or habit of thought that the philosopher Arthur Lovejoy (1933) labeled the principle of plenitude. [For a discussion of the economic implications of this principle, see Warsh (1984).] This principle states that the world is full: every conceivable entity already exists. As a corollary, it follows that nothing truly new can ever come into being. Every conceptual possibility already has a realization in the physical world. To an economist, it means that we can always assume that we are in the interior of the goods space.

As Lovejoy shows, the principle of plenitude is fully formed in the writings of Plato and has played a central role in Western thought ever since, appearing prominently, for example, in the writings of Spinoza and Leibniz. The set of all conceivables entities corresponds to Plato's world of ideals. For every ideal there must exist a corresponding entity in the world of experience, for otherwise the world would be incomplete and imperfect.

Much of the thrust of modern scientific inquiry has been directed at overcoming our innate prejudice in favor of predetermination and plenitude. Many educated people are contemptuous of creationists who claim that there is a supreme being that created all existing forms of life according to a master plan and who deny the possibility that different forms of life emerged by chance and will continue to do so. We also ridicule the turn of the century head of the patent office who recommended the abolition of the patent system because everything had already been invented. Yet in their everyday approach to economic problems, most economists adopt a modeling strategy that reflects an implicit belief closely related to that of the patent clerk and the creationist.

The durability of the principle of plenitude can presumably be traced to a deeply rooted human desire to understand the world and therefore to believe that the world is capable of being understood. If we admit that new things can happen – that there are many things that could exist that do not yet exist – we undermine our most common explanation of why the world is the way it is: It has to be this way for it could not have turned out otherwise. For example, if there is only one conceivable form of intelligent animal, then we can explain why humans had to emerge from evolution in precisely the form that they did. People simply could not have turned out differently.

When it is applied in a specific context such as evolution, it is now obvious that the principle of plenitude is not just false; it is wildly misleading. The fossil record shows us that there are many conceivable types of animals that once existed but no longer exist. There are an extraordinarily large number of types of forms of life that could havé evolved but did not because of a variety of historical accidents. If the earth had not been hit by the meteor that killed the dinosaurs, the forms of life on earth today would be different in ways that we can hardly begin to imagine. Scientifically, a far better guiding principle would be that of sparsity: only a vanishingly small fraction of all conceivable entities can actually exist in the physical world.

When we are confronted with things we do not understand, we all retreat into a world view that has more in common with the creationists than we would like to admit. Take cosmology. Implicitly, most of us believe that there is only one kind of physical universe that is conceivable. We can therefore explain (in some limited sense) why the universe is the way it is. It could not have turned out otherwise. For example, space must have three dimensions instead of two or four because, we assert, there is no such thing as two or four dimensional real space, even *in principle*. As unsatisfying as this style of explanation may be, most of us find it preferable to the alternative. It is deeply unsettling to admit that humans, other animals, our planet, even the universe itself, are merely the result of a long sequence of accidents that determined which of many different conceivable outcomes were actually realized.

Modern cosmology suggests our universe itself is new. At some date there was a big bang when both space and time as we know them came into existence. And as if this were not enough, some physicists are now seriously raising the possibility that it was only our particular universe that was created in what we refer to as 'the' big bang. This extension of the inflation model of big bangs suggests that our universe emerged from a small local fluctuation in the energy density in some prior universe, and that many comparable universes are continually branching off from our own and many other universes, expanding in their own big bang. Because these other universes start from slightly different initial conditions, the physical laws that they obey may be radically different from our own.

We would prefer to believe that there is only one kind of universe and that any universe could not possibly have turned out differently from the one we inhabit. But to theoretical physicists, the humorous physics exam question, 'Define universe and give two examples', is starting to look serious. Perhaps the fact that space in our universe is three dimensional is an accidental outcome that is no more explicable than the fact that there are no mammals that lay eggs and fly.

Whatever one's reaction is to the possibility that there could be, or will be, or are many different universes, nothing could be more certain than the facts that there are many different types of economic goods, that there will be many more of them in the future, and that these will somehow be selected from an incomprehensibly larger set of goods that could conceivably be produced. To see why the set of all conceivable goods is so much larger than the set of goods that could ever be produced in our universe, one need only do a few simple combinatorial calculations. Consider for example, the set of all possible computer programs that could fit on the lowest capacity floppy disk still in use for personal computers. Such a floppy disk can store a bit string consisting of 360,000 × 8 positions, each of which contains either a 0 or a 1. This means that there are about $2^{3,000,000}$ or about $10^{1,000,000}$ different conceivable computer programs (i.e. bit strings) that could fit on such a disk. This is a big number. For comparison, about 10^{18} seconds have passed since the big bang started the universe (or the current version of our particular universe). It is estimated that there are about 10^{100} electrons and protons in the visible portion of our universe.

Most of the possible bit strings will be totally useless as computer programs, but unless we are willing to side with the patent clerk and suggest that all the useful software has already been written, there must be a small fraction of these possible programs, perhaps just one in every $10^{999,900}$, that will do wonderful things. This would still leave 10^{100} interesting programs to discover. Given the physical constraints we face it is inevitable that the vast majority of these useful programs will never be written or tested. For example, even if we could use every elementary particle in the universe to code a different bit (spin up could stand for 1, spin down for 0), the total storage capacity of our universe would be 10^{100} bits. Since each useful program would require far more than one bit to store its code, there would not be enough capacity in the entire universe to store the code of all the useful programs. It is an inevitable fact of life that economies will forever operate on the boundary of goods space, that only a small subset of all possible goods will ever be introduced.

The principle of plenitude manifests itself in the everyday operation of economic analysis in the implicit assumption that nevertheless, all the relevant goods already exist. It lets us think of economic analysis as taking place in the interior of goods space. It lets us tell our students that the essential economic problem is depicted in fig. 1, not fig. 3. Economists do recognize that the set of traded goods in an economy is always changing, but according to the economic version of plenitude, this turbulence is an epiphenomenon of no fundamental interest. According to this view, you can change the labels on a box of detergent, but detergent is just detergent. According to this view, all decisions in private markets can accurately be characterized as being like changes in the allocations of the quantities of a fixed and unvarying set of underlying goods that cover all of the relevant possibilities. Perhaps the clearest statement of this point of view comes in the characteristics space descriptions of goods articulated by Lancaster (1966). All apparently new goods are just different bundles of a fixed underlying set of primitive goods.

Economists have not always gone this far. Schumpeter was quite explicit about the central importance of the creation of genuinely new goods (1934). So was Young (1928). Yet in our post-WWII enthusiasm for distilling the 'miracle of the market' down to its mathematical essence, economists have generally been willing to push these issues aside. Decentralized markets could be shown to get everything right only by assuming that half of our basic economic problem (and by far more difficult half at that) had already been solved. In the Arrow–Debreu world, all the relevant or useful dated and state-contingent goods have already been selected from the large set consisting of all possible goods. The only problem that remained for the economic system to solve was to allocate existing goods between a fixed set of existing uses.

It was perhaps this step, the implicit incorporation of the principle of plenitude, rather than the unification of statics and dynamics per se, that troubled those economists who dissented from the general equilibrium approach to economic theory. The refutation of the principle of plenitude seems to lie at the heart of the case that has been made for path dependence and the role of history in economic analysis by Arthur (1989) and David (1985).

Much of the explicit discussion by the critics of general equilibrium theory has focused on issues other than path dependence or the sparsity of the set of realized goods. Instead, critics have emphasized the suboptimality of market outcomes and the implausibility of the assumption of rational choice. On these issues, the mainstream economists and the critics are not very far apart. Mainstream economics is filled with arguments showing that equilibria are in many cases not first-best Pareto optimal, and mainstream equilibrium theorists increasingly emphasize the importance of informational constraints, costs of decision making, and of the importance of learning processes that converge slowly, if at all, to the outcome derived from the assumptions of fully informed rational calculation. Disagreements between the critics and mainstream economists over these issues are really over no more than differences in degree.

It is a deeper difference in the presumptions that different people have about the nature of the world and the nature of the problem that an economic system must solve that separates the different groups of economists and makes communication so difficult. When economists insist on the importance of disequilibrium behavior, part of what they must be saying is that out of all conceivable goods, only a very sparse subset can actually exist in any real economy, so genuinely new goods can always be added. Any analysis that treats a dynamic economy as being formally equivalent to a static economy characterized by plenitude – fullness in the set of goods – cannot, according to this view, capture the essential aspects of growth and change.

After being lectured about the philosophical origins underlying the selection of different areas in fig. 4, the economist subjected to the Rorschach test described in the previous section could complain of having been tricked. If we had explained that the demand curve in fig. 4 was for a good that did not yet exist, then he would of course have commented on the welfare significance of the triangle *A*. Moreover, this economist could observe with irritation, the point argued here (about newness) is ‘not new’. There is an old literature on marginal cost pricing and utility regulation [Hotelling (1938)] that is concerned precisely with the problem of paying the fixed cost of bringing some service into being. Similarly, economists have long understood that the problem of assigning patents or copyrights involves a fundamental conflict between conflicting efficiency conditions. Monopoly profits induce people to incur the fixed costs of invention, but cause distortions in the allocation of goods once the invention exists. We have always known, this economist could continue, that if the market had to solve two problems – both introducing the right goods and allocating them properly – a competitive economy would never be able to get the prices right. (Section 6 below challenges the fall-back position that is sometimes offered suggesting that an economy characterized by monopoly power can get the prices right if price discrimination is tolerated.)

But the question remains, why do economists assume, unless instructed otherwise, that all of the relevant goods already exist? Why is it that the problems associated with the introduction and creation of goods are always part of the footnotes or the marginalia? Why is it that we devote so much more attention to the problem illustrated in fig. 1 than to the one illustrated in fig. 3?

5. Building a bridge

To illustrate how attention to the creation of new goods can change the emphasis one places on different aspects of economic analysis and to emphasize that the arguments raised here are indeed quite old, this section recapitulates the first analysis of the problem of bringing a new good into existence. As head of an engineering district in 19th century France, Jules Dupuit was responsible for building roads, bridges, and canals. Because the tools he needed for cost–benefit or project analysis did not exist, he invented them. In a paper published in 1844, he described both the demand curve and the revenue (i.e. 'Laffer') curve. He developed the familiar triangle approximation to the welfare cost of a tax, and noted that it varied with the square of the tax wedge. Apparently without knowledge of the work of Cournot, he developed the theory of monopoly pricing. To this he added an analysis of the welfare benefits of price discrimination.

As an engineer and a civil servant, Dupuit was motivated by the practical problem of discovering rules he could use to decide whether a specific investment project like a bridge should be built. Among his many cogent observations was the remark that the maximum net revenue that could be collected by simple monopoly pricing is less than the total social value created by the bridge. Thus, even if the firm that builds the bridge has perfect property rights and can extract the entire monopoly revenue forever, it collects an amount that is strictly less than the value of the bridge to society.

The monopolist can capture the rectangle $B+D$ in fig. 4 but not the surplus triangle A. In honor of his observation and to distinguish it from the usual deadweight triangle C, it is appropriate to call triangle A a Dupuit triangle.

Harold Hotelling reprised Dupuit's analysis in his 1938 paper on railway and utility pricing, and a substantial debate over marginal cost pricing ensued. [See Ruggles (1949–1950) for a survey.] The discussion followed exactly the lines one would predict on the basis of the principle of plenitude. Economists like Hotelling recognized that there is a fixed cost that must be covered if the bridge builder is to break even, but all of the attention is addressed to the formal question of how to avoid the losses associated with deadweight triangles. With few exceptions [e.g. Coase (1946)], the analysis ignores Dupuit's practical concern, how to decide which of the many conceivable bridges in the world should actually come into existence. Economists like Hotelling implicitly took it as given that we knew which bridges to build (and which ones not to) and worried only about how to finance them.

To the modern reader, the description given by Hotelling or Meade (1944) of the advantages of having the government build bridges, operate railroads, and socialize production in all other instances where there are large fixed costs seems rather naive. In part, this reflects subsequent experience with how governments actually function. We have seen too many cases in which a bridge or roadway is built at great expense and then allowed to deteriorate beyond repair because the government in a developing nation or a city in the United States postpones relatively minor maintenance expenditures. But our reaction to the analysis is also colored by a sense that the welfare losses under consideration are relatively small. Few modern economists think that the deadweight triangles associated with bridge tolls or railroad fares that exceed marginal cost add up to welfare losses that are more than a trivial fraction of total GDP.

Our sense of the relative magnitudes derives in large part from a pioneering set of papers published in the 1950s by Arnold Harberger. They offered back-of-the-envelop calculations of the economy-wide welfare losses caused by resource misallocations. The first paper focused on monopoly power distortions in the United States analogous to those caused by setting a bridge toll that is higher than marginal cost [Harberger (1954)]. Total losses, Harberger calculated, were on the order of a few tenths of a percent of GNP. The second paper estimated the costs of trade restrictions for the Chilean economy: about 2.5% of GDP [Harberger (1959)]. These numbers had the bracing effect of forcing economists to be more explicit about the relative importance that should be attached to each of the many conceivable distortions that could exist in an economy. It was no longer enough to put a sign on a welfare effect. The magnitude matters as well.

The key result behind such calculations was noted by Dupuit: welfare distortions depend on the square of the wedge between price and marginal cost. This observation is typically offered as a justification for why government intervention to remove small distortions is not very important and why private sector distortions like monopoly power are not very costly. But it also implies that the costs of misguided, harmful government intervention will also be relatively small. Viewed from this perspective, the textbook description of market equilibrium offers up a remarkably weak defense of laissez faire policy. If markets are perfectly competitive, government intervention can do no good, but it also does little harm.

Yet as noted above, the current professional consensus (mirrored for example by the World Bank's *World Development Report* for 1991) is that the policies of protection and import substitution followed for many years by countries like Chile severely reduced welfare. There are only two ways to reconcile this view with the logic behind Harberger's estimates. The estimates of the losses from deadweight triangles might be too low because the true size of the traded goods sector, the true elasticities of demand, or the true tariff rates are an order of magnitude larger than the ones used in his calculations, but we can measure these quantities relatively well and an order of magnitude error seems quite unlikely. A more likely explanation is that the losses associated with deadweight triangles may indeed be small, but that there are other, much more important losses associated with the loss of large numbers of Dupuit triangles.

To illustrate how the Dupuit triangles might enter the analysis, return to Dupuit's analysis of a bridge. Suppose that a foreign firm would be willing to incur a fixed c_0 to build a bridge if it could charge the simple monopoly price for bridge crossings. Consider the pricing problem for the bridge builder under the simplifying assumption that the marginal cost of bridge crossings in zero. If the bridge is built, the Dupuit triangle under the demand curve and above the monopoly price represents a pure surplus gain to society. The deadweight triangle represents the additional gains that could be achieved if the price for crossings were set equal to the marginal cost of zero.

Let $p(x)$ denote the inverse demand curve for bridge crossings. A monopolistic bridge builder will maximize revenue $p(x)x$. The familiar first-order condition is

$$p(x)+p'(x)x=0. \tag{1}$$

In this simple one-period model, assume that the revenue that the monopolist collects is larger than the cost c_0 of building the bridge. If it were not, the firm would not be willing to undertake construction. (In a multiperiod model, c_0 would be the interest and maintenance cost of the bridge in each period.)

Now suppose that the government levies a simple ad valorem tax τ on bridge crossings. The monopolist maximizes net revenue $(1-\tau)p(x)x$. The new first-order condition is

$$(1-\tau)p(x)+(1-\tau)p'(x)x=0, \tag{2}$$

which clearly yields the same solution for x and p as does the monopoly problem with no taxes. Provided that the bridge is still built, the tax acts like a pure profit tax on the bridge builder (which we can take to be a foreign owned multinational corporation, the profits of which are of no concern to the domestic government). The tax has no effect on the size of the deadweight triangle and causes no additional distortions. So far, we have only restated the usual result that a pure profits tax induces no distortions.

But at the point where the after-tax revenue just equals the cost of the bridge, $(1-\tau)p(x)x=c_0$, the welfare effects of the tax change discontinuously. For any larger tax, the bridge is not built and the entire Dupuit surplus triangle is lost to the economy. Because the tax extracts wealth from foreigners, domestic welfare increases monotonically with the tax rate, up to the point where the bridge is not built. Then welfare drops sharply, to a level below the welfare achieved when no tax is imposed.

For a developing country, far more is at stake than bridges, roads and the kind of civil engineering project that we usually associate with cost–benefit analysis. The more important losses are likely to arise from the absence of economic goods that are never imported and made available in a developing country – the capital goods that are not imported, the production processes that could be but are not undertaken there, the many possible entrepreneurial activities that are never attempted.

One of the fundamental premises of the argument offered here is that bringing each of these different kinds of activities or goods into existence requires a fixed cost expenditure. For a developing country, multinational firms are one of the most important sources of new economic activities, and the fixed costs they face are easy to imagine if one works through the details of their operations. One can think, for example, of the information a foreign retailer must collect about quality, reliability, and capacity of suppliers before it can begin to buy garments assembled in a new country. The retailer would have to establish new financial relationships for clearing transactions and new shipping and communications links for moving goods. It would have to learn about the local legal, regulatory, and tax environment, and it would have to investigate the nature of political risk. It would also need to invest in long term implicit and explicit contractual relationships for the trading relationship to be successful.

Evidence that fixed costs are important comes from the observation that many services and goods are simply not available at any price in many parts of the world. If there were no fixed costs, one should find that all possible goods, services, production processes and types of exchange are available to firms located everywhere in the world. If local taxes or other distortions are high or if transportation cost were high, prices might be higher there, but the goods would still be available. One need not, however, rely purely on indirect evidence. Direct studies have shown, for example, that it can be quite costly to transfer production technology from an advanced to a developing country [Teece (1977)].

For the rich class of goods, services and activities contemplated here, there are many kinds of impediments that play the function of the tax rate τ in the bridge example. Tariff and non-tariff barriers to the imports of key inputs and components, excessive tax rates, restrictions on ownership, domestic content requirements, corruption, and bureaucratic regulation can all have the same effect as τ, reducing the revenue stream that acts as an inducement for someone contemplating an initial fixed cost investment. The cumulative effect of these indirect costs can be very large. A test case conducted in Peru in 1983 demonstrated that it took four university students 289 days of full time work to get the 11 permits needed to legally open a small garment assembly shop [de Soto (1989)]. In addition to reducing profits for any new enterprise, these kinds of distortions may also cause the usual deadweight losses, but as Harberger's analysis shows, the deadweight losses are unlikely to explain the large welfare costs that we now attribute to extensive intervention.

In his analysis of prospects for Chile, Harberger pointed to the dynamic process of technological change and diffusion of knowledge as the source from which large welfare gains could be extracted. If we interpret the process of diffusion and adoption of new technologies in a developing country as resulting from the introduction of new economic activities that are made possible by fixed cost investments of the kind noted above, then Harberger's diagnosis is exactly right. The true costs of badly designed government interventions, and especially of trade restrictions in developing countries, come not from their effects on the static allocation of resources between the activities in an economy that already exist. Rather, they come from the stifling effect that the distortions have on the adoption of new technologies, the provision of new types of services, the exploitation of new productive activities, and on imports of new types of capital goods and produced inputs.

If one implicitly relies on the principle of plenitude and assumes that all of the relevant or possible productive activities already exist in a developing country – that is, if one assumes that the essence of economic development is just to do more of the things that the economy already does – Harberger's analysis shows that the costs of realistic distortions cannot be too large. But if once one recognizes that there are many kinds of inputs in production, many kinds of productive activities, many kinds of expertise and insight from the rest of the world that could be but are not in use in a developing economy, one sees that the welfare costs from distortions can be much higher.

The textbook analysis may therefore have it backwards. It may be taxes on ex post profits, not taxes that change relative prices, that have the most severe welfare consequences. If we restrict attention to the allocation of resources between a fixed set of goods, taxes on profits are indeed irrelevant. But if the decisions about which goods to introduce are the most important economic deicisions that an economy makes, profit taxes may be the ones that truly matter. Moreover, an ex post profit tax that prevents a good or activity from coming into being is particularly insidious because it leaves no trace of what might have been. The bridge that is not built is as easy to overlook as the dog that did not bark.

6. Price discrimination

As Dupuit observed in his partial equilibrium analysis, perfect price discrimination would solve the problem of financing the fixed cost of introducing a new good if it were feasible. Decentralized market mechanisms would then be able to solve the problem of selecting which goods to introduce, and equilibria would be Pareto optimal. A modified version of the price system, with a price schedule or a multi-part pricing arrangement, would be able to solve both the problems facing an economy.

This section shows that two factors undermine a reliance on price discrimination to achieve the social optimum. The first problem is that complete surplus extraction has distributional effects that may be a serious policy concern in a developing nation. The second, and conceptually more important point illustrated here is that multi-part pricing and price discrimination cannot be supported in a full equilibrium in many situations of practical interest. In these cases, the partial equilibrium intuition derived from the study of a single demand curve does not apply.

The infeasibility of a Pareto optimal decentralized equilibrium with multi-part pricing is central to the specific claim of this paper – that trade restrictions cause first-order welfare losses because the decentralized equilibria that actually arise are not first-best Pareto optimal. But the analysis in this section is also intended as an illustration of the positive contribution that formal mathematics can make to economic analysis. As noted in the introduction, it is likely that an emphasis on explicit mathematical formalism impeded the early analysis of many issues in economics because we initially relied on assumptions like convexity that were very restrictive. The example in this section shows that when we learn to weaken those restrictive assumptions, the formal mathematics can reveal points that informal verbal arguments miss.

Imagine that a small, less developed, economy produced output using capital K and labor L according to a Cobb–Douglas production function $Y=L^{1-\alpha}K^{\alpha}$. Suppose that a foreign industrialist has invented a new kind of capital input x. The new capital good can then be used with labor and old capital to produce output according to $Y=L^{1-\alpha}(K^{\alpha}+x^{\alpha})$. This functional form is a simple special case of the constant elasticity of substitution aggregator of the capital goods K and x: $(K^{\rho}+x^{\rho})^{\alpha/\rho}$. If ρ is equal to 1, K and x are perfect substitutes and final production takes the usual Cobb–Douglas form in terms of labor and a single capital aggregate. Imagine, however, that in this example, K represents structures and x represents machinery. Because x is a new kind of capital good that is not a perfect substitute for existing capital goods, ρ must be different from 1. For simplicity, let ρ equal α, so that K and x enter production in an additively separable fashion. By setting ρ bigger than or less than α, one could make K and x into complements or imperfect substitutes.

The relevant question in this simple example is whether the foreign industrialist can use price discrimination or multi-part pricing to capture all of the benefit associated with the introduction of the good x into this economy. Assuming that the industrialist has strong property rights over the invention of the good x – or equivalently, that the technology for producing x is secret – the industrialist can certainly charge the simple monopoly price for units of x used in this economy. The derived demand for x from competitive firms that produce Y according to the specified technology is

$$p(x)=\alpha L^{1-\alpha}x^{\alpha-1}. \tag{3}$$

From this expression, it is clear that no matter what level of x the industrialist chooses to supply, simple monopoly pricing lets him capture only a fraction α of the increase in output that his actions induce. For any selected level of x, output goes up by $L^{1-\alpha}x^{\alpha}$, but the industrialist captures only $p(x)x=\alpha L^{1-\alpha}x^{\alpha}$. It is tempting to assume that the firm that purchases x captures the Dupuit surplus triangle and that the industrialist could prevent this by setting a price schedule or by charging a fixed fee to each buyer plus a charge for each unit sold. Assume for the moment that the industrialist has property rights that are so strong that he can prevent the secondary market sales that typically undermine these mechanisms for extracting surplus. The more fundamental problem here is that the buyers of the capital goods have no surplus that the industrialist can extract. They are competitive firms. They operate under conditions of constant returns to scale. They always earn zero profits.

If the industrialist did structure pricing in an attempt at extracting more value than he could extract under simple monopoly pricing, there would be no equilibrium in this economy. Prior to the introduction of x, let w be the wage for labor and let r be the rent on the existing capital good K. Suppose that the industrialist offers a firm a pricing schedule that extracts a total payment $P(x)$ for x units of the new good. Suppose that the industrialist chooses this schedule so that it extracts all of the additional output that the introduction of x induces. Before x is introduced, Euler's theorem tells us that the firm produces output Y equal to $wL+rK$. It earns zero profit. After being supplied with x units of the new good, the firm would produce output equal to $wL+rK+P(x)$ since by assumption, $P(x)$ is equal to the increase in output induced by x. The firm would still break even if wages for labor and the rental rate on capital K remained the same after x is introduced. If only a single small firm were offered the use of x, wages would indeed remain constant. But of course, a profit maximizing industrialist will want to sell to all firms that produce final output. In this case, the market wage for labor will increase to a new value $w'>w$ because the new units of x raise the marginal productivity of labor. And if wages go up, all final output firms will now earn negative profits, and the pricing schedule $P(x)$ cannot be sustained.

If the industrialist tries to scale back the pricing schedule to keep his customers from failing, he will be forced all the way back to a constant price per unit of the good x. To see why, let q denote the price of the last unit of good x purchased under the price schedule. The output producing firm will purchase units of x until its marginal productivity is equal to q. By Euler's theorem, if w' is equal to the new marginal productivity of labor and if r is equal to the marginal productivity of capital, the value of total output for a price-taking firm that employs K units of structures, L units of labor, and x units of machinery will be $Y=w'L+rK+qx$. Consequently, if the industrialist charges a price per unit q for all of the units of machinery, the final output firms just break even. Any attempt to extract additional surplus by imposing either an access fee for the right to buy the new capital good or a higher price for the initial units of the good x will fail.

If the industrialist captures only a fraction of the benefits from the introduction of the good x and if the producers of output earn zero profits, the remainder of the surplus benefits must accrue to one of the other two factors of production in this economy, old capital or labor. Because K and x are additively separable, the introduction of x has no effect on rent for K, $r=\alpha L^{1-\alpha}K^{\alpha}$. Labor captures all of the surplus. After good x is introduced, total wage income increases by $(1-\alpha)L^{1-\alpha}x^{\alpha}$, and this amount plus the revenue captured by the industrialist is just equal to the increase in total output.

In this kind of general equilibrium setting, price discrimination does not help a monopolist extract surplus, but vertical integration can, at least in some circumstances. Suppose that the industrialist refuses to sell his good x. Instead, he purchases K and L and uses them together with x to produce output directly. By doing so, he can take over all production of final output and keep wages at their previous level, $w=(1-\alpha)L^{1-\alpha}K^{\alpha}$. In this way, he can avoid the wage gains that would otherwise give a fraction of the increase in output to labor. This pattern of industrialization without wage gains is what it would take to ensure that the industrialist captures all of the benefits he creates when he introduces machinery. It is not a prescription for development that the citizens of a developing country will want to encourage. (As an aside, note that it also cannot be a historically accurate description of the process of development in industrialized countries, for it if were, unskilled labor would still earn what it earned prior to the industrial revolution.)

In addition to its problematic implications for the distribution of the gains from industrialization, this kind of vertical integration can support the Pareto optimum only if there is no other sector in the economy that uses labor. Suppose, for example, that it is also possible to produce agricultural output using land T and labor according to a standard, concave production function $F(T, L)$. As x is introduced into the economy, the marginal product of labor in the production of the manufactured good Y will increase, so optimality requires that labor shift from agriculture into manufacturing. The industrialist can act as a monopsonist, inducing labor into the manufacturing sector and raising wages only to the extent that it is profitable for him to do so, but this will not lead to the Pareto optimum for the standard reasons. And contrary to the usual partial equilibrium intuition, once again there is no way for the industrialist to do better with price discrimination. Any scheme that draws labor out of agriculture will raise the marginal productivity of labor in agriculture, and an industrialist who controls the entire manufacturing sector can do nothing to stop competition from agriculture from raising the wages he must pay. Only if the industrialist can integrate horizontally and bring the entire economy under his control will he be able to shift labor between sectors without changing wages. Only a command economy can move resources between sectors without changing prices. The defense of monopoly therefore degenerates into the traditional defense of central planning.

Brown et al. (1992) give conditions under which equilibria with multi-part pricing will exist and a positive quantity of the good controlled by a monopolist will be supplied. The reason why their result does not apply here is revealing. Their analysis assumes that monopolists sell directly to final consumers. They show that if the total value that the consumers place on the good in question is greater than the costs to the monopolist, an equilibrium with multi-part pricing and positive supply of the good will exist. Their result depends, however, on the presence of sufficient 'willingness to pay' on the part of the people who purchase directly from the monopolist. The examples in this section show that if the customers a monopolist sells to are competitive firms or if the inputs a monopsonist purchases are also used by a competitive industry, the willingness to pay will be zero. Competitive firms that buy the new input have no willingness to pay for the introduction of the new good because they earn zero profits regardless of whether the good is introduced or not. An individual laborer has no willingness to pay for the introduction of the new good even though it increases wages for all workers; if workers in the industrial sector had to chip in to pay for the introduction of the new good, they could simply move to the agricultural sector and get the benefits of higher wages without having to help pay for the new goods.

Even though the aggregate benefits from the introduction of the good exceed the cost of introducing it, the benefits do not accrue to the competitive firms that are the buyers or to the specific inputs that the monopsonist directly employs. As a result, multi-part pricing can not be sustained unless the competitive markets that shift the benefits throughout the economy are completely shut down.

7. Relative magnitudes

The previous section shows that simple monopoly pricing and Dupuit surplus triangles are endemic to decentralized economies. They will be present whenever the firms that introduce new goods draw inputs from competitive industries or sell their output to competitive industries. Since these Dupuit surpluses form the basis of the wage gains that we associate with development, it is a good thing for the citizens of the developing world that they exist.

The analysis so far has shown that any intervention that prevents a new activity from coming into existence will be bad for development. It remains to show that aggregate losses from lost Dupuit triangles can be large enough to be of serious policy concern.

To see why it is possible in principle for trade distortions to impose large social costs, look again at the areas A and C under the demand curve in fig. 4. Dupuit's analysis taught us that the welfare losses associated with deadweight triangles are second-order small, and Harberger's analysis showed that the aggregate losses from deadweight triangles typically amount to a small fraction of GDP. But Dupuit triangles can be quite large even when the wedge between price and marginal cost is small. In fact, they get larger as the wedge gets smaller. As a result, welfare losses vary in direct proportion with the number of Dupuit triangles that are destroyed or prevented from coming into existence. A rough guide to the welfare losses in any country will therefore be the difference between the range of productive inputs that are available there and the range of productive inputs that could be put to use there.

To illustrate these issues in a formal model, suppose that output can be written as a function of labor L and a large quantity of different types of capital goods indexed by i:

$$Y=L^{1-\alpha}\sum_{i=1}^{N}x_i^{\alpha}. \tag{4}$$

This is just a many-good version of the additively separable, constant elasticity aggregator for a large number of different kinds of capital goods. Instead of thinking of just two broad categories such as structures and machinery, let i index many different types of capital goods – blast furnaces, lathes, fork lift trucks, looms, etc. – that are not perfect substitutes for each other.

As before, this functional form imposes the restriction that output must be a constant returns to scale function of L and all N of the different inputs x_i. To maintain the parallel with the model in the last section, all of the x's will be referred to as capital goods. They could, of course, represent a much broader class of inputs in production, including intermediate inputs or imported forms of specialized human capital. As in the example from the last section, labor's share of total income will be $1-\alpha$, and in the aggregate, the share of all the capital goods will be α.

To close the model, it remains to specify how the different capital goods x_i can be produced and how N, the number of different types of goods is determined in equilibrium. If this were intended as a model of growth in an advanced economy, the right way to close the model would be to specify a research and development technology for inventing new goods. [Different ways of doing this are illustrated in Romer (1990), Rivera-Batiz and Romer (1991), and in the closely related model in Grossman and Helpman (1991).] But for a developing economy, it is more appropriate to assume that a very large set of productive inputs already exists in the rest of the world, and each of them can be introduced into this economy after incurring a cost of $c_0(i)$ that can depend on the index i. Instead of the single foreign entrepreneur described in the last section who can introduce a new technology, there are a large number of different foreign entrepreneurs, each of whom can incur a cost and introduce a new type of productive input into this economy.

For example, if good 27 represents a rubber-tired front-end loader, the fixed cost $c_0(27)$ can be interpreted as the fixed cost of setting up a service and parts supply network necessary before these loaders can be used in this economy. Alternatively, good 27 could be the services of an engineering consulting firm that helps manufacturing firms implement quality control systems in production, and $c_0(27)$ would represent the fixed costs of setting up a local branch consulting office. The goods are arranged so that c_0 is increasing in i. For simplicity, we can assume that this dependence is linear, $c_0(i)=\mu i$. We can also choose the units for measuring quantities of the x_i's so that c_1, the marginal cost of one additional unit of each good, is the same for all goods.

Eq. (4) describes the constant returns to scale production function that competitive firms use to produce final output from labor and all available capital goods x_i. For each possible input in production, there is a foreign entrepreneur that contemplates paying the fixed cost $c_0(i)$ of entering the local market. If it enters, it will maximize profit subject to the downward sloping derived demand curve from the final goods producers. If the ex post monopoly revenue it can extract is greater than $c_0(i)$, firm i will enter. Good N is the marginal good, the one which entry costs just equal ex post monopoly revenue.

This setup is of course just a many-good version of the bridge problem outlined in section 5. Also, for the reasons outlined in section 6, only simple monopoly pricing can be sustained in equilibrium because the many different suppliers of capital goods all sell to competitive firms.

Because of the symmetry between all of the capital goods as inputs in final production, all the firms that enter face the same derived demand and earn the same revenue. Differentiating output with respect to the quantity of good i gives the marginal productivity schedule or industry inverse demand curve, which takes exactly the same form as in the previous example.

$$p_i(x_i)=\alpha L^{1-\alpha}x_i^{\alpha-1}. \tag{5}$$

Because we want to study the effects of distortions in this economy, suppose that the government imposes an ad valorem tax or tariff τ on all of the purchases of foreign imported goods. If the firm selling good i enters, it faces a profit maximization problem of the form

$$\max_x \, (1-\tau)p_i(x)x-c_1x. \tag{6}$$

Because the demand curve is a constant elasticity demand and because the cost of each unit of the capital good made by firm i is constant, it is easy to verify that the solution to this problem takes the familiar form of a constant proportional markup, $p_i^*(\tau)=c_1/\alpha(1-\tau)$. Monopoly revenue net of cost and taxes is then equal to $c_1x^*(1-\alpha)/\alpha$. In equilibrium, the input level $x^*(\tau)$ will be the same for all goods and can be found by putting the expression for p^* on the left side of eq. (5) to yield

$$x^*(\tau)=\alpha^{2/(1-\alpha)}c_1^{-1/(1-\alpha)}(1-\tau)^{1/(1-\alpha)}L. \tag{7}$$

The expressions for p^* and $x^*(\tau)$ can then be substituted into the expressions for the ex post net revenue. Ignoring integer constraints, the solution for the equilibrium number of inputs $N(\tau)$ comes from equating the fixed costs of introducing the marginal good, μN, to the expression for net revenue:

$$N(\tau)=\frac{1-\alpha}{\alpha\mu}c_1x^*(\tau)$$

$$=\frac{1-\alpha}{\mu}\alpha^{(1+\alpha)/(1-\alpha)}c_1^{-\alpha/(1-\alpha)}(1-\tau)^{1/(1-\alpha)}L. \tag{8}$$

Using eq. (4) and the symmetry of all of the inputs, gross domestic product $Y_{\mathrm{DOM}}(\tau)$ can be written in terms of N and x:

$$Y_{\mathrm{DOM}}(\tau)=L^{1-\alpha}N(\tau)[x^*(\tau)]^{\alpha}. \tag{9}$$

Because all of the capital inputs are purchased from abroad, gross national product (that is, income for the citizens of the domestic economy) Y_{NAT} is equal to labor's share of gross domestic product plus the tax revenue collected by the government. Because the total payment by domestic firms for capital inputs is equal to a fraction α of $Y_{\mathrm{DOM}}(\tau)$ and because the government collects a fraction τ of these payments as tax revenue, national income can be written as

$$\begin{aligned}Y_{\mathrm{NAT}}(\tau)&=(1-\alpha)\,Y_{\mathrm{DOM}}(\tau)+\tau\alpha\,Y_{\mathrm{DOM}}(\tau)\\&=(1-\alpha+\tau\alpha)\,Y_{\mathrm{DOM}}(\tau).\end{aligned} \tag{10}$$

To calibrate this simple economy, let $\alpha=0.5$. This is a rough compromise between the need to keep the markup of price over marginal cost, $1/\alpha$, from being too large and to keep the share of labor, α, from being too small. It also makes the algebra easy.

There are two tax experiments that one can conduct in this simple economy. The first is to calculate the losses in output that would occur if the government were to impose an unexpected tariff (or tax) τ after firms have already made their entry decisions. Because entry costs are sunk costs, this kind of unanticipated tax will have no effect on the number of goods N that are available for use in this economy. This kind of calculation is implicitly what economists do whenever they hold the set of goods constant and consider only the deadweight losses from a tax or tariff distortion. Formally, this amounts to inserting $N(0)$ in place of $N(\tau)$ in the expression for $Y_{\mathrm{DOM}}(\tau)$. A simple calculation shows that in this special example, the efficiency loss for this economy, measured as a fraction of national income in the absence of a tariff varies exactly with the square of the tax or tariff rate:

$$\begin{aligned}1-\left.\frac{Y_{\mathrm{NAT}}(\tau)}{Y_{\mathrm{NAT}}(0)}\right|_{N=N(0)}&=1-\frac{1-\alpha+\tau\alpha}{1-\alpha}\frac{N(0)}{N(0)}\left(\frac{x^*(\tau)}{x^*(0)}\right)^{\alpha}\\&=1-\frac{(1-\alpha+\tau\alpha)(1-\tau)^{\alpha/(1-\alpha)}}{1-\alpha}\\&=\tau^2.\end{aligned} \tag{11}$$

The last equality holds because $\alpha = 0.05$.

The other experiment is to consider the effects of a fully anticipated tax. In this case, the tax reduces N in addition to reducing x. Firms that contemplate entry understand that the monopoly revenue collected after entry will be smaller because of the tax. In this case, the expression for the proportional loss in output increases much more rapidly with τ:

$$1 - \frac{Y_{\mathrm{NAT}}(\tau)}{Y_{\mathrm{NAT}}(0)} = 1 - \frac{1-\alpha+\tau\alpha}{1-\alpha}\,\frac{N(\tau)}{N(0)}\left(\frac{x^*(\tau)}{x^*(0)}\right)^{\alpha}$$

$$= 1 - \frac{(1-\alpha+\tau\alpha)(1-\tau)^{(1+\alpha)/(1-\alpha)}}{1-\alpha}$$

$$= 2\tau - 2\tau^3 + \tau^4. \qquad (12)$$

The difference between the welfare losses implied by these two different experiments is striking. If the tariff rate τ is 10%, the first calculation – the one that holds constant the set of goods in use – implies that national income falls by only 1%, the square of the tax or tariff rate. The second calculation – the one that lets the set of good vary in response to reductions in anticipated revenue – implies that national income falls by 19.81%. And if $\tau = 0.25$, the first calculation implies that national income falls by 6.25%. The second implies it falls by about 47%. And depending on how one values government revenue, the effect could be even worse. Gross domestic product and labor income fall by about 58% when the tax rate is 25%, but government receives revenue equal to about 11% of the no tax GDP. If part of this tariff is dissipated by collection costs and rent seeking, the loss to society could be closer to the 58% fall in GDP than to the 47% fall in national income.

There are aspects of this calculation that could bias the estimates of the welfare losses in either direction. Setting α equal to 0.5 implies that the elasticity of demand for the capital goods is 2, a value that is too high and leads to an overstatement of the welfare losses in both experiments. On the other hand, the assumption that the effects that the different types of capital goods have on output are additively separable could understate the losses when the number of goods N can vary. Many different kinds of capital goods and many different kinds of economic activities will actually be complements, so that a reduction in the number of activities and goods that are available locally will have an additional damping influence on the incentive to undertake any particular activity that this calculation does not pick up. Recognizing that there are these strong complementarities can help explain why economic activity tends to cluster in particular geographical regions and why policy changes can sometimes induce a shift from stagnation to rapid 'take off', In particular, complementarities help explain why direct foreign investment in developing nations – the primary means whereby productive activities and equipment are transplanted from OECD countries – tends to be so heavily concentrated in just a handful of countries. [For evidence of the importance of direct foreign investment in the process of development and of its sensitivity to the costs of doing business, see, for example, Romer (1993).]

It is also important to remember that this is an example, not a general proof. The simple results that emerge in this case depend on the choice of simple functional forms and a convenient choice of the parameter α. A single example is sufficient, however, to overturn the widespread presumption that the costs of trade restrictions are always small. In at least some circumstances, the difference between holding the set of goods constant and allowing it to vary increases the estimated efficiency losses by an order of magnitude.

8. Conclusion

Showing that something can be true in a model does not make it so. Only evidence can settle an assertion of fact such as the one made here – that trade restrictions, taxes, corruption, bureaucratic red tape, and the many other small contributions to the cost of doing business in a developing country can have very large negative effects on aggregate output because they can sharply reduce the number of productive activities that are undertaken there.

In addition, even if we agree that the model is correct, it is not a simple matter to translate the model's insights into predictions or policy implications about the world. The model and thought experiment considered here examine the effects of tariffs on imports. For simplicity, the discussion does not consider the possibility that foreign producers can undertake production in the developing country and avoid the tariffs. In practice, it is conceivable that some protectionist policies induced this kind of tariff-jumping investment for large developing countries and that this process allowed, or even encouraged, increases in the domestic availability of some goods even when tariff barriers were high. The general failure of import substitution policies suggests that the costs of tariff barriers eventually outweigh any benefits that come from tariff-jumping, but to understand why import substitution policies seemed to work in the early years after they were adopted (at least in some countries), one would have to include domestic production in the analysis. In this extension of the basic model, the general lesson from the analysis will carry over. What matters for the workers in a developing country are the other inputs in production that are available locally for use with their labor. The costs to these workers of implicit and explicit taxes on the firms that provide new goods and services may be very high.

Even though it is not easy to verify a model is right, and even if it is difficult to apply the insights from a simple model in a complicated world, it is important that policy makers and economists be willing to consider models that allow for new goods. Models matter because they shape the point of view that we adopt, and our point of view directly influences how we process, interpret, store, and recall the large quantity of evidence that is available to us. Formal theoretical analysis can contribute to our use of evidence largely by forcing us to make explicit the habits of mind that we take for granted. Mathematical models help us examine our implicit assumptions more critically, explore their implications more systematically, and consider alternatives to them more freely. As evidence that exposure to models matters for our reading of the evidence, observe that many intelligent people who are not trained in economics live their whole lives surrounded by evidence that demand curves slope down, yet never recognize or understand this simple fact. People trained in the basics of the supply-demand model see the evidence quite differently.

The fundamental claim in this paper is that economists, particularly economists trained in formal mathematics and general equilibrium theory in the last four decades, have adopted a point of view and used a collection of models that has led them to substantially underestimate the aggregate effects of additions to the cost of doing business. In particular, they have underestimated the costs of tariffs and other restrictions on international trade. Economists who were familiar with the experience of particular countries have often asserted that price distortions, taxes, tariffs, and bureaucratic impediments could be of decisive importance in slowing development, but they often made these claims without much in the way of theoretical back-up, and without convincing important parts of the profession. The accepted theoretical approach strongly suggested that distortions of reasonable size should have only small negative effects of aggregate output, and evidence alone is usually not enough to convince people. It takes a new theory (or at least a new point of view) to beat an old one.

The point of view that impeded progress in economics – one asserting that nothing new ever comes around – is deeply ingrained in human thought and is not easy to overcome. But the convergence of efforts by economists working on international trade, economic growth, path dependence in economic history, and innovation at the level of the firm or industry may finally generate enough focussed attention on this mental stumbling block to make some headway in overcoming it.

If so, new growth theory will have earned its name, but not in the literal sense in which it has come to be used. As this paper has emphasized, many of the specific claims now being emphasized in growth theory have roots that go back at least 150 years. The larger issues go all the way back to Plato. So there is little that is truly new about the theory. Nor should the name be interpreted in the ironic sense (emphasized by quotation marks placed around the word new) in which it was first used by critics. There is nothing to recommend the pursuit of novelty for its own sake, and no one seriously devoted to the study of growth would argue that originality is what makes this work important.

The term has a different spin that was entirely unplanned but is nonetheless apt. New growth theory may not *be* new, but it is *about* newness. And newness, like history, matters.

References

Aghion, P. and P. Howitt, 1990, A model of growth through creative destruction, Econometrica 60, 323-351.

Arthur, W. Brian, 1989, Competing technologies, increasing returns, and lock-in by historical events, Economic Journal 99, 116–131.

Brown, Donald J., Walter P. Heller and Ross M. Starr, 1992, Two-part marginal cost pricing equilibria: Existence and efficiency, Journal of Economic Theory 57, 52–72.

Coase, Ronald H., 1946, The marginal cost controversy, Economica 13, Reprinted in: The firm, the market, and the law (University of Chicago Press, Chicago, IL, 1988).

David, Paul, 1985, Clio and the economics of qwerty, American Economic Review 75, 332–337.

de Soto, Hernando, 1989, The other path (Harper and Row, New York).

Dixit, A. and J. Stiglitz, 1977, Monopolistic competition and optimum product diversity, American Economic Review 76, 297–308.

Dosi, Giovanni, Christopher Freeman, Richard Nelson and Richard Silverberg, 1988, Technological change and economic theory (Pinter Publishers, London).

Dupuit, Jules, 1969, On the measurement of the utility of public works, in: Kenneth J. Arrow and Tibor Scitovsky, ed., Reprinted in Readings in welfare economics (Irwin, Homewood, IL).

Feenstra, Robert C., 1992, How costly is protectionism, Journal of Economic Perspectives 6, 159–178.

Grossman, G. and E. Helpman, 1991, Quality ladders in the theory of growth, Review of Economic Studies 58, 43–61.

Grossman, G. and E. Helpman, 1992, Innovation and growth in the global economy (MIT Press, Cambridge, MA).

Harberger, Arnold C., 1954, Monopoly and resource allocation, American Economic Review 44, 77–87.

Harberger, Arnold C., 1959, Using the resources at hand more effectively, American Economic Review 49, 134–146.

Helpman, Elhanan and Paul Krugman, 1985, Market structure and foreign trade (MIT Press, Cambridge, MA).

Hotelling, Harold, 1938, The general welfare in relation to problems of taxation and of railway and utility rates, Econometrica 6, 242–269.

Jones, Lawrence and Rodolfo Manuelli, 1990, A convex model of equilibrium growth: Theory and policy implications, Journal of Political Economy 98, 1008–1038.

Krugman, Paul, 1990, Rethinking international trade (MIT Press, Cambridge, MA).

Lancaster, Kelvin, 1966, A new approach to consumer theory, Journal of Political Economy 74, 132–157.

Lovejoy, Arthur O, 1933, The great chain of being (Harvard University Press, Cambridge, MA).

Lucas, Robert E., Jr., 1988, On the mechanics of economic development, Journal of Monetary Economics 22, 3–42.

Meade, James E., 1944, Price and output policy of state enterprise: A symposium, The Economic Journal 54, 321–339.

Nelson, Richard and Sydney Winter, 1982, An evolutionary theory of economic change (Harvard University Press, Cambridge, MA).

Rebelo, Sergio, 1991, Long run policy analysis and long run growth, Journal of Political Economy 99, 500–521.

Rivera-Batiz, Luis and Paul M. Romer, 1991, Economic integration and endogenous growth, Quarterly Journal of Economics 106, 531–556.

Romer, Paul M., 1986, Increasing returns and long-run growth, Journal of Political Economy 94, 1002–1037.

Romer, Paul M., 1987, Growth based on increasing returns due to specialization, American Economic Review 77, 56–62.

Romer, Paul M., 1990, Endogenous technological change, Journal of Political Economy 98, S71–S102.

Romer, Paul M., 1991, Increasing returns and new developments in the theory of growth, in: William Barnett, Bernard Cornet, Claude d'Aspremont, Jean Gabszewicz and Andreu Mas-Colell, eds., International symposium in economic theory and econometrics (Cambridge University Press, Cambridge).

Romer, Paul M., 1993, Two strategies for economic development: Using ideas and producing ideas, Proceedings of the Annual World Bank Conference on Development.

Ruggles, Nancy, 1949–1950, Recent developments in the theory of marginal cost pricing, Review of Economic Studies 17, 107–126.

Schultz, Theodore W., 1975, The value of the ability to deal with disequilibria, Journal of Economic Literature 13, 827–846.

Schumpeter, Joseph, 1934, The theory of economic development (Harvard University Press, Cambridge, MA).

Segerstrom, Paul, T.C.A. Anat and Elias Dinopoulos, 1990, A Schumpeterian model of the product life cycle, American Economic Review 80, 1077–1091.

Teece, David, 1977, Technology transfer by multinational firms: The resource cost of transferring technological knowhow, Economic Journal 87, 242–261.

Warsh, David, 1984, The idea of economic complexity (Viking Press, New York).

Young, Allyn, A., 1928, Increasing returns and economic progress, Economic Journal 38, 527–542.