Chapter 24

INCREASING RETURNS, IMPERFECT COMPETITION AND THE POSITIVE THEORY OF INTERNATIONAL TRADE

PAUL KRUGMAN*

MIT, Cambridge

Contents

I would like to thank Don Davis and Gene Grossman for extremely helpful comments. Don Davis, in addition to providing invaluable help in revising an early draft, did the yeoman service of presenting that draft in my absence.

Handbook of International Economics, vol. III, Edited by G. Grossman and K. Rogoff

0. Introduction

This chapter is an awkward one to write, because it is in effect squeezed between an illustrious predecessor and the topics covered in other chapters. On one side, I need not retrace the ground covered by Elhanan Helpman's ground-breaking survey in the original *Handbook of International Economics* (written in 1982). On the other side, it is probably fair to say that most of the innovative work on increasing returns and imperfect competition in international trade theory since the late 1980s has focussed on dynamic issues, especially technological change – and these issues are covered in Chapter 2 and other chapters in this volume.

Why, then, write this chapter? For three reasons. First, Helpman's survey turns out to have come a few years too soon. Perhaps inevitably given the state of the field at that time, the general impression conveyed in that chapter was of a collection of highly disparate and messy approaches, standing both in contrast and in opposition to the impressive unity and clarity of constant-returns, perfect-competition trade theory. And yet within only a few years after Helpman's survey, it had become clear (largely due to his own work) that the new ideas were not such a grab-bag after all. On the contrary, many of the insights of increasing returns trade theory could be understood in terms of a quite simple common framework, in which trade has the effect of moving the world economy toward the "integrated economy" that would exist if national boundaries could be eliminated. The integrated-economy approach also suggested considerable continuity between the new elements in trade theory and the older tradition. Furthermore, the framework provided a "grammar" that could be used to discuss topics that went well beyond the standard analysis of trade in final goods, such as the effects of trade in intermediate goods and the role of multinational firms in the world economy.

One purpose of this chapter, then, is to present a compact restatement of this integrated economy approach. In effect, the first part of the chapter is a *Reader's Digest* version of Helpman and Krugman (1985).

At the same time, since Helpman wrote his survey there have been a number of developments in the theory of trade (and direct investment) under imperfect competition that cannot be represented using the integrated economy approach, typically because they rely on some kind of persistent market segmentation that trade does not fully eliminate. These developments include, most notably, analyses of the effects of the size of the domestic market, and of price discrimination; they also have a bearing on recent thinking about multinational enterprise. So a second purpose of this chapter is to summarize some of these post-1982 developments.

Finally, the years that have passed since the original Helpman survey offer us a chance to gain some perspective. Now that the "new trade theory" has grown

middle-aged along with its founders, we can try to ask what it accomplished and what it left undone.

This chapter, then, is in three parts: a restatement of the integrated-economy approach to trade theory, a survey of other developments that cannot be treated within that approach, and a brief reconsideration of the achievements and limits of the now not-so-new "new trade theory".

1. The integrated-economy approach to international trade

1.1. Samuelson's angel

It is common in expositions of international trade theory to start by imagining two isolated countries, which then are allowed to begin trade with each other. The integrated economy approach goes in the reverse direction, starting from a unified economy, then breaking it up.

Perhaps the first suggestion of using this approach to think about international trade came in a parable used by Paul Samuelson to explain the concept of factor price equalization. Here's the parable: Once upon a time, all the factors of production in the world were part of a single economy, able to work freely with each other. This integrated world economy had reached an equilibrium, with all the things that go with such an equilibrium: goods prices, factor prices, resource allocations, and so on.

And then down came an angel. (Although Samuelson does not say so, this is obviously the angel from the Tower of Babel story; presumably the factors of production had dared to challenge heaven, and were being punished for their presumption.) The angel smote each unit of each factor of production on the forehead, labelling it as belonging to a particular nation; and thenceforth factors could only work with other factors from the same country.

But how much damage had the angel done? Well, perhaps none. Provided that the angel had not divided the factors of production too unevenly between the nations, it might still be possible through specialization and trade to achieve exactly the same global output and consumption as before. In that case, trade would have the effect of "reproducing the international economy"; and one could indeed describe such a restoration of the integrated economy (which would involve, among other things, equalization of factor prices) as the purpose of international trade.

Samuelson's angel story can be given a very convenient representation in a two-factor, two-good, two-country, constant returns world. Consider Figure 1.1, which was originally suggested by several authors in the 1960s but whose dissemination is due to Dixit and Norman (1980). The box in Figure 1.1 represents the resources of the world economy as a whole, with the height of the box representing the world supply of capital, the width the world supply of labor. In the pre-angel, integrated economy there will be full employment of both factors; we represent the

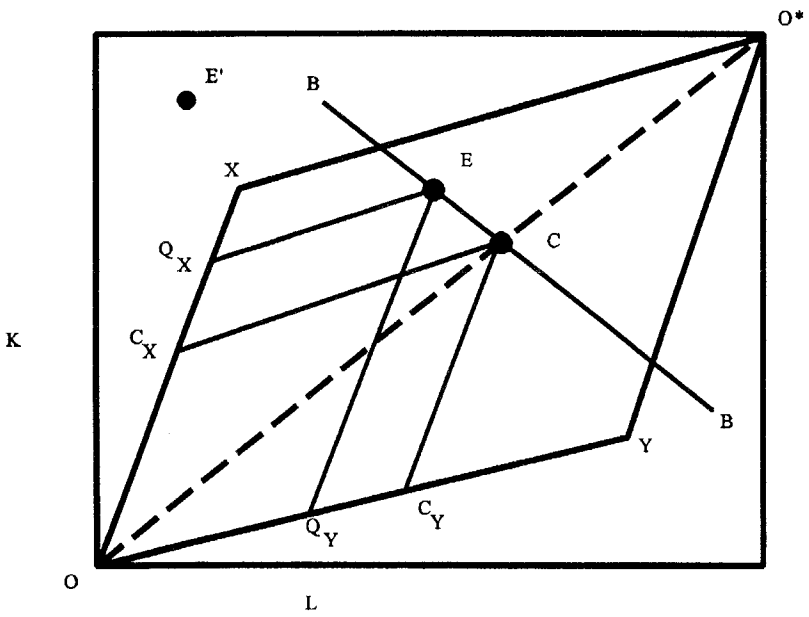

Figure 1.1.

resources devoted to the capital-intensive industry by OX, those devoted to the labor-intensive industry by OY. The two vectors must, of course, sum to the world factor endowment – that is, the completed parallelogram just fits into the box.

Now let the angel come down, and divide the factors of production into two nationalities, Home and Foreign. Let us measure Home endowments from the southwest corner of the box, O, and Foreign endowments from the northeast corner, O^*. Then the angel's punishment can be represented by a single point which, measured from O, shows the Home endowment, and which shows the Foreign endowment measured from O^*. Point E is one example of a world endowment point; E' is another.

Did the angel do any harm? In order to reproduce the integrated-economy outcome, it must be possible to use the same production techniques (i.e. capital–labor ratios) that were used in the integrated economy to produce the same outputs, while fully employing the resources of each country. It is immediately obvious that this is possible if the endowment point is, like point E, within the parallelogram OXO^*Y. Indeed, we can determine the required resource allocation simply by completing parallelograms. Let Home devote resources OQ_X to production of X, OQ_Y to production of Y; and let Foreign correspondingly devote resources Q_XX and Q_YY to the two industries. That will precisely replicate the integrated economy production, while fully employing both countries' resources. It is also immediately apparent that

this is an equilibrium, one in which both countries have the same factor prices as in the integrated economy (and therefore the same as each other). In this case, then, the angel's action was a warning, not an actual punishment.

It is conversely clear that if the endowment point lies outside the parallelogram, at a point such as E', there is no way to replicate the integrated economy. Among other things, factor prices will therefore not be equalized. It then becomes a fairly nasty business to determine the details of the pattern of specialization.

Returning to the case in which the integrated economy is reproduced, we have seen how the pattern of resource allocation and production is determined. Can anything more be said? Let us make one more assumption: that everyone has the same homothetic preferences. Then everyone in the economy will consume the two goods in the same ratio. But this means that the factor services embodied in each individual's consumption must also be in the same ratio – and this ratio must be the same as the world ratio of capital to labor. Geometrically, the factor services embodied in each country's consumption must lie on the diagonal OO^*.

Where on this diagonal does the consumption point lie? At that point at which income equals spending for each country. But all income takes the form of factor earnings. Suppose that E is the endowment point. Then draw a line (BB) with a slope $-w/r$, where w is the wage rate and r the rental rate on capital in the integrated equilibrium, passing through E. All combinations of factor services on that line have the same value, and the factors embodied in Home consumption must therefore be at the point where BB crosses the diagonal, C.

Once we have determined this consumption point, several things become obvious. First, since E represents the factor services embodied in Home production and C the factors embodied in Home consumption, EC is Home's net trade in factor services. Second, we can read off Home's consumption of each good (or more precisely, the resources used to produce its consumption of each good) by completing parallelograms: OC_X is consumption of X, OC_Y is consumption of Y. And therefore the pattern of trade in goods can also be seen: Home exports $OQ_X - OC_X$, imports $OC_Y - OQ_Y$.

The end result, of course, is to give yet another statement of the Heckscher–Ohlin theorem that the capital-abundant country Home exports the capital-intensive good X. Here, however, we view the theorem in a different light: the point of trade in this case is to reproduce the integrated economy by trading embodied factor services; trade in commodities is simply a means to that end.

One convenient aspect of this diagram is that it can readily be adapted to situations in which there are more than two goods. Suppose, for example, that there are three goods. Then the situation is as represented in Figure 1.2. Here we let OX be the vector of resources that the integrated economy would employ in producing the most capital-intensive good, XY the resources in the good of intermediate capital intensity, and YO^* the resources in the most labor-intensive good. Clearly, the integrated economy can be reproduced as long as the endowment point lies within the hexagonal area traced out by these vectors. The actual pattern of production and trade is

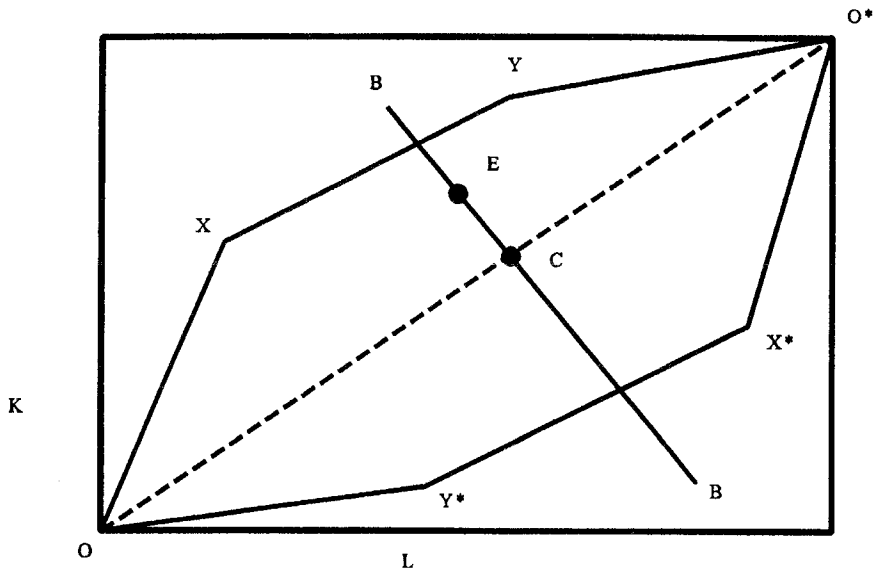

Figure 1.2.

indeterminate, since there are many ways to fully employ the resources of each country. However, the Heckscher–Ohlin principle is still honored in the sense that the net trade in embodied factor services must still be *EC*. That is, the capital-abundant country is a net exporter of the embodied services of capital, a net importer of the services of labor.

The most useful thing about this approach, however, is that it can be applied to models in which returns to scale are not constant, and in which differences in factor proportions are not the only motive for trade.

1.2. Differentiated products

During the course of the 1970s several industrial organization theorists, notably Dixit and Stiglitz (1977) and Lancaster (1975) offered ways to give the old tradition of Chamberlinian monopolistic competition solid microfoundations, and to embed monopolistic competition in differentiated products within general equilibrium models. These new models depended on very special and essentially implausible assumptions about tastes and technology, but they were quickly recognized as useful devices for thinking about a variety of issues involving increasing returns and imperfect competition. In particular, towards the end of that decade several international trade theorists realized that the new models of monopolistic competition

offered a particularly clean way to think about trade based on economies of scale rather than comparative advantage. These initial developments, and the issues associated with the choice of alternative ways of modeling product differentiation, were covered in the 1982 Helpman survey and need not be reviewed here.

At first it seemed that the various models of trade with monopolistic competition were essentially inconsistent both with the more traditional factor-proportions approach and with each other. I think that it is fair to say that this would have appeared to be the case even to most readers of Helpman's survey, although Helpman (1981) made a major effort to unify the new approaches with the older tradition. By thinking in terms of the integrated-economy approach, however, it becomes easy to embed trade in differentiated products within a factor proportions model. Indeed, one does not even need to redraw the diagram!

Let's change the story of Samuelson's angel slightly. We now imagine that one of the industries in the pre-angel world – say industry X – was characterized by Chamberlinian monopolistic competition. That is, there were many firms in that industry, each a little monopolist producing a distinct product with a technology that exhibited internal economies of scale – but free entry into the industry had driven firms to the tangency position in which no economic profits were being earned. (The zero-profit assumption implies that all income is still factor income.) We need not specify the details of how products were differentiated – whether individuals loved variety, as in Dixit–Stiglitz models, or whether each individual consumed a preferred variety, as in Lancastrian models – except to assume that all varieties were symmetric.

Now let the angel come down, and divide the world into nations. What must be done to reproduce the integrated economy? As before, the pre-angel output of each industry must be allocated among countries in such a way as to fully employ all factors of production using pre-angel techniques. But what does it mean to allocate the production of an industry that consists of many differentiated products? It could mean allocating a certain share of the production of each variety to each country. This, however, would not reproduce the integrated economy, because production of each variety would take place at a smaller, and thus less efficient scale. To reproduce the integrated economy, then, one must allocate each *variety* of X to only one country. If there are 100 X varieties, and Home gets 57 percent of the world X production, then it must produce all of the pre-angel output of 57 varieties, while Foreign produces the other 43. (It is indeterminate which varieties are produced in which country, but it doesn't matter, because the varieties are by assumption symmetric.)

We can still use Figure 1.1 to describe the aggregate outcome. If the endowment point is E, Home will still devote resources OQ_X to the production of X, and consume a volume of X products that required OC_X to produce. We must now, however, reinterpret the quantity $OQ_X - OC_X$. It no longer represents Home's total exports of X. The reason is that X is now not a homogeneous good but a group of differentiated

products, and the varieties produced in Foreign are different from those in Home. As a result, Home consumers will spend some of their income on X-varieties produced in Foreign (either because individuals have a taste for variety, or because there is a dispersion of tastes). That is, Home will import as well as export X. We must therefore reinterpret $OQ_X - OC_X$ as Home's *net* exports of X.

The overall pattern of trade can be schematically represented by Figure 1.3, where the length of arrows represents the value of trade. Home is a net exporter of the monopolistically competitive good X and an importer of Y; this pattern of *interindustry* trade can be viewed as the comparative advantage component of the trade flow. However, there is additional trade over and above this interindustry trade, because Home and Foreign produce different varieties of X; this two-way trade within the X industry constitutes the *intraindustry* component of international trade.

The essential reason for this intraindustry trade is the existence of economies of scale. That is, if varieties of X were produced under constant returns, it would be possible to reproduce the integrated economy by dividing the production of each variety between the countries, and there would be no need to engage in intraindustry trade to satisfy consumers' taste for variety; it is increasing returns that prevent each country from producing the full range. Thus in this model we can make a simple distinction: interindustry trade – which allows countries to trade embodied factor services – is the result of comparative advantage; intraindustry trade – which allows countries to retain the integrated economy scale of production – is the result of increasing returns.

As one might expect, the relative importance of these two kinds of trade depends on the difference in the countries' factor abundances. Imagine sliding the endowment point E back and forth along the line BB (which amounts to varying the resources of the countries while keeping their relative economic size constant). If E is moved down to the diagonal OO^*, there will be no net trade in factor services, and hence no interindustry trade. However, trade will not vanish, because the two countries will

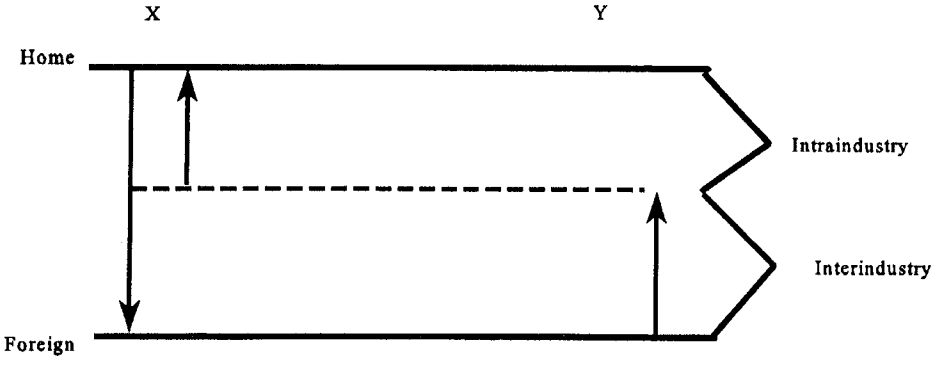

Figure 1.3.

still be producing different varieties of X. Thus all trade will be intraindustry. Conversely, move E to the edge of the parallelogram. Then Foreign will produce no X, and there will be only interindustry trade. So countries with similar resources will tend to engage primarily in intraindustry trade, countries with very different resources will tend to engage primarily in interindustry trade.

By using the integrated economy approach, then, we are able to accomplish several things. First, we can show that the details of particular monopolistic competition models of trade, such as the form of product differentiation, do not matter when it comes to the description of the motives for and pattern of trade. Second, we can very easily offer a picture of trade in which both economies of scale and comparative advantage are motives for trade. Finally, this approach makes the role of increasing returns seem less a departure from, than a natural extension of the grand tradition in trade theory: both trade in embodied factors and specialization to achieve scale economies are ways in which the trading world "tries" to reproduce the integrated economy.

1.3. External economies

Until the monopolistic competition models of trade emerged in the late 1970s, nearly all thinking about increasing returns in trade took the form of models in which economies of scale were purely external to firms. While the tradition of such models had some influence, however, it was limited, probably for three reasons. First, the disembodied, abstract nature of pure external economies may have seemed unsatisfactory: how would you recognize such external effects if you saw them? Second, external-economy models of trade typically seemed to yield a bewildering variety of equilibria, leaving the modeler with a taxonomy rather than a clear set of insights. Finally, much of the traditional literature on external-economy models regarded the effect of external economies as being one of modifying or distorting the pattern of specialization away from that implied by resource abundance; given the dominance of comparative advantage thinking, most international economists instinctively tended to regard any such effects as probably minor.

The 1982 Helpman survey offered a major clarification of the role of external economies in trade – indeed, it offered an interpretation of equilibrium in an externality-ridden world that was later to be rediscovered with considerable fanfare in the "new growth" literature [Romer (1986)]. Nonetheless, the integrated economy approach allows us to add some further insights.

External economies are by no means always, or even usually, consistent with the integrated-economy approach. There are, however, some cases in which external economies are consistent with that approach, and those cases do offer some interesting intuition into the possible role of external economies in the world economy.

1.3.1. The pattern of trade

Consider an integrated economy that produces *three* goods, one of which is subject to industry-specific external economies. That is, production of X takes place under constant returns from the point of view of the individual firm, but the efficiency of each atomistic firm is an increasing function of total industry output. Despite these external economies, there will normally be a competitive equilibrium in this integrated economy. (If the external effects are very strong, there may be multiple equilibria even in the integrated economy; we disregard this possibility.) Associated with this integrated economy equilibrium will be resource allocations to all three industries as well as goods and factor prices.

Now once again we invoke Samuelson's angel. As before, the angel divides factors between two nations. He or she must also, however, make a critical decision about the extent of the punishment: do external benefits spill over between the nations? If there is full international spillover of externalities – if factors of production in each country derive the same external benefits from X production abroad as at home – then all that is needed to reproduce the external benefits of the integrated economy is that total world X production be the same as integrated economy X production – which is the same criterion that applies in the absence of externalities. That is, factor price equalization and trade can be represented as in Figure 1.2, with the presence of external economies making no difference as long as E lies within the hexagon.

But suppose that the angel's rules are stricter, and that external benefits accrue only from production that takes place in the same country. Then the integrated-economy outcome can be reproduced only if we add another criterion to those of matching integrated economy resource allocation and full employment: *all of the industry that is subject to external economies must be concentrated in one country.*

Figure 1.4 shows what this criterion implies. In the figure, OX represents the resources that the integrated economy would have devoted to the production of the good subject to external economies; XY and YO^* represent the resources devoted to two constant-returns sectors. (It is, of course, not necessary that the external-economy sector be the most capital-intensive; it is straightforward to see how the diagram changes if it is of intermediate capital intensity.) In order to reproduce the integrated economy, we must concentrate the X industry either in Home or in Foreign; the other two industries can then be allocated between the countries to fully employ their resources. This means, then, that the integrated economy can be reproduced as long as the endowment point lies inside either of the solid parallelograms in the figure. (These parallelograms are shown here as overlapping, but they need not be.)

The choice of words here is deliberate: provided that E lies in the right region, the integrated economy *can* be reproduced. That is, there then exists an equilibrium that reproduces it. There may also be other equilibria which do not. This point is easier to make in a two-good, one factor model: it is a familiar result there that there may be both an equal-wage equilibrium in which the increasing returns industry is concen-

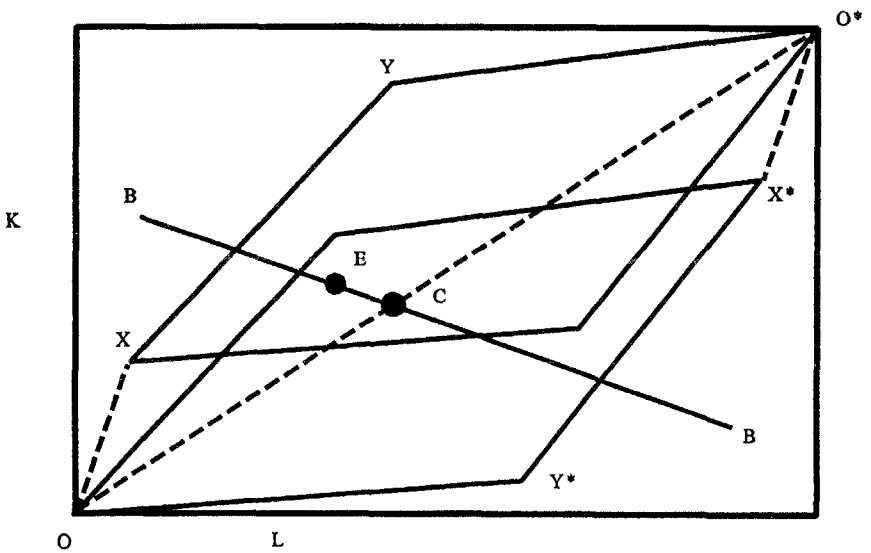

Figure 1.4.

trated in the larger country, and an unequal-wage equilibrium in which the smaller country is specialized in the increasing-returns sector [see, for example, the analysis in Ethier (1982a)]. The former equilibrium reproduces the integrated economy, while the latter does not. Thus it is something of a cheat in this case to focus only on equilibria that reproduce integrated economy; but we will nonetheless do so.

Suppose, then, that the endowment point lies within one or both of the parallelograms, and that we consider only equilibria that reproduce the integrated economy. We can then say several useful things about international trade.

First, even if the countries have identical factor abundances – if E lies on the diagonal line OO^* – X production will be concentrated in only one country, and there will therefore be international trade. Thus in an external-economy model, as in the monopolistic competition models, international trade will occur as a result of increasing returns even if there is no comparative advantage.

Second, the pattern of production and trade may be indeterminate. Suppose that E, as drawn, lies in the overlap between the two parallelograms. Then there are two patterns of production (and hence trade) that reproduce the integrated economy: one in which X is concentrated in Home, one in which it is concentrated in Foreign.

Third, although there may be indeterminacy, the factor proportions theory continues to hold in the sense that the capital-abundant country is a net exporter of capital services embodied in goods. In Figure 1.4, as in earlier figures, EC represents the net trade in factor services. Thus this approach suggests the happy interpretation that increasing returns add an overlay of additional specialization and trade to the

trade required to embody the necessary movement of factor services, adding to comparative advantage rather than modifying or distorting it (although we no longer have the simple distinction between intra- and inter-industry trade).

1.3.2. The gains from trade

Much of the traditional literature on external economies in trade seemed to imply that such external economies created a conflict of interest between countries; that there could be a more or less zero-sum battle over who got the external benefits of increasing-returns industries. And even in Helpman's 1982 survey he seemed to suggest that welfare might decline if your country saw its increasing-returns sectors shrink as a result of trade. The particular version of external-economy trade described here suggests, however, an alternative and more benign view.

First, even when the pattern of trade is indeterminate – where either country might end up with the X industry – as long as the integrated economy is reproduced it doesn't matter for either country's welfare, since the owners of factors in either case are exactly as well off as they were pre-angel.

But are countries better off with trade than they would be in its absence? It is possible to establish a sufficient criterion for such gains: a country gains from trade if the *world* output of the external-economy good with trade is larger than that country's *individual* output of that good would be in the absence of trade. This seems a fairly weak test, and therefore suggests a more optimistic view about the welfare effects of trade under external economies than the older literature.

The criterion may be justified by an argument developed in Helpman and Krugman (1985), and illustrated in Figure 1.5. In that figure, the curve II represents the unit isoquant for some good before trade; we suppose that the initial factor prices are represented by ww, so that V represents the initial unit factor inputs.

The opening of trade will have two effects. First, factor prices may change. Second, if this good is subject to external effects, the unit isoquant may shift. Suppose first that the good is not subject to external effects. Then the iso-value line of the inputs used to produce a unit of the good after trade will be a line with a slope different from ww, say $w'w'$. Obviously, V must lie above that line. That means, however, that the *post-trade* income of the inputs used to produce the *pre-trade* output of a good is always more than enough to purchase that pre-trade output. But if this is true for all goods, then post-trade income is always more than enough to purchase pre-trade consumption – which means that the country's choice has been expanded, and that it gains from trade.

If a good is subject to external effects, the unit isoquant may shift. What matters for the price of the good is the isoquant *in the country in which the good is produced.* As long as that isoquant shifts inward – as in the illustrated isoquant $I'I'$ – then V is still certain to lie above the iso-value line, which is now shown as $w''w''$. That is, it remains true, indeed is true a fortiori, that after trade the pre-trade inputs can more

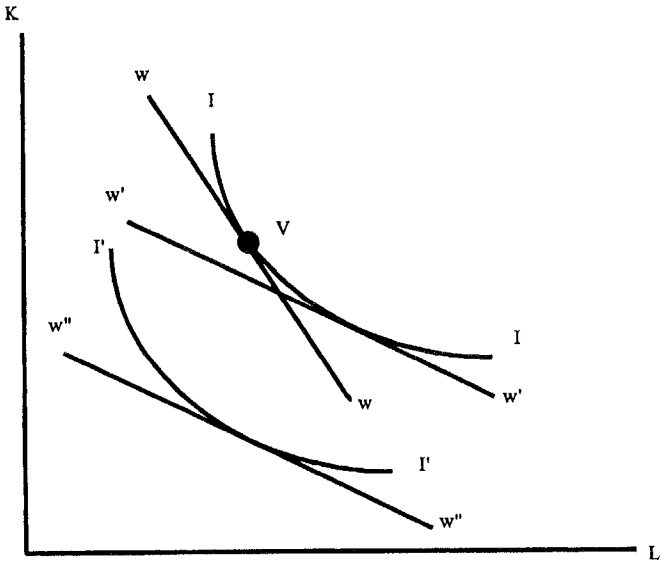

Figure 1.5.

than afford to buy their previous production. But the unit isoquant for the producing country will indeed be inside our autarky isoquant as long as that country's output is larger than ours would have been in the absence of trade.

We are therefore able to justify our criterion for gains from trade: a country gains as long as the *world* output of each external-economy good, wherever it may be located, exceeds our autarky output. The gains from trade, like trade itself, arise both from comparative advantage – reflected here in a change in factor prices – and from the additional external economies achieved through geographic concentration of industries, reflected in the inward shift in the unit isoquant.

All of this suggests a rather benign view of trade under external economies. According to this view, it is a good thing for the world that there be a Silicon Valley, that is, that the semiconductor industry be concentrated *somewhere*; it really doesn't matter where. The conclusion is, of course, sensitive to the model, but it at least serves as a corrective to the assumption that country-specific external economies always imply a struggle over who gets the good sectors.

1.4. Intermediate goods, traded and nontraded

In some of the early papers on increasing returns in trade, Ethier (1979, 1982b) argued that intraindustry trade in practice consists largely of trade not in differentiated

final goods but in differentiated inputs, and that the possibility of such trade in effect gives rise to international (as opposed to national) external economies. As we have just seen, even in the case of purely national external economies the welfare gains from these external effects may be effectively globalized. Nonetheless, Ethier's point appears to be largely if not completely correct about actual intraindustry trade, and in any case the implications of trade in differentiated intermediate goods – or its absence – are a subject of considerable interest.

Imagine a world economy in which there are two sectors X and Y, and in which X is a monopolistically competitive, differentiated-product industry. The varieties of X do not, however, enter into final consumption. Instead, they are used as inputs into the production of Y.

What effect does this have on the model? If both goods are costlessly tradeable, the fact that one of the industries is an intermediate good makes little difference. Indeed, we can still apply Figure 1.1: trade will reproduce the integrated economy as long as the endowment point lies within the parallelogram, and the volumes of both intra- and inter-industry trade will be exactly as in the case of two final goods.

Matters become very different, however, if differentiated intermediate goods are made nontradeable.

Nontraded final goods can be introduced into the integrated-economy framework with little difficulty. Essentially, factor-price equalization can still occur for a non-trivial set of endowments as long as there are at least as many tradeable sectors as factors of production. The region of endowments leading to factor price equalization is, however, smaller the larger the share of expenditure on nontradeables.

The case of nontraded intermediates poses additional issues. Suppose that an upstream industry provides many differentiated inputs to a downstream sector; that each of those inputs is produced subject to scale economies; and that the inputs cannot be traded (or indeed have transport costs). Then the integrated economy can only be reproduced if *all* of the intermediate varieties and the good whose production uses them are concentrated in the same country! This obviously rules out reproducing the integrated economy in the two-industry case just described.

We can still make use of the integrated-economy approach, however, if we introduce some extra industries. Suppose, in particular, that there are *four* industries: N, X, Y, Z. Industry N produces differentiated and nontradeable inputs used in the production of X; Y and Z are constant-returns. In Figure 1.6, we let ON, NX, XY, and YO^* represent the integrated-economy vectors of resource use in each industry. (Again, the factor-intensity rankings are arbitrary.)

In order to reproduce the integrated economy, the "industrial complex" consisting of the N and X industries must be concentrated in one country. The Y and Z industries can then be allocated between the countries. On reflection, the implication is apparent: trade can reproduce the integrated economy as long as the endowment point lies either in the parallelogram defined by Y and Z after putting the N–X complex in Home, or in that defined after placing N–X in Foreign. Thus the region of such endowment points is the shaded area in the figure.

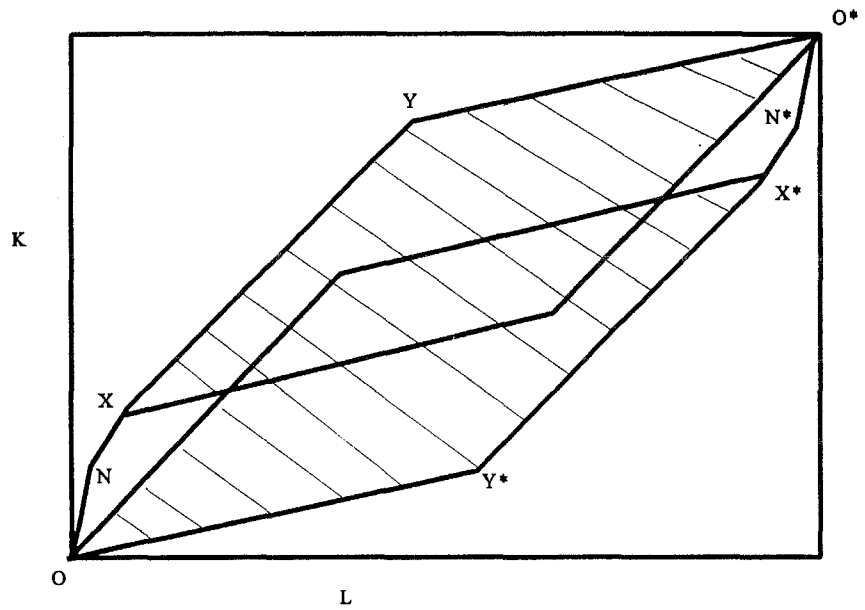

Figure 1.6.

This looks familiar. It is, in fact, essentially the same result as that shown in Figure 1.4, where the X industry was subject to pure national external economies. All of the other implications of that model carry over as well: trade due to both increasing returns and comparative advantage, determinate net trade in factor services but possibly indeterminate trade in goods, and (with some care) the proposition that all countries benefit from the establishment of an industrial concentration somewhere, no matter where. The reassuring implication is that the insights from external economy stories continue to apply when the externalities result from market-size effects rather than purely disembodied spillovers.

1.5. Multinational enterprise

A final application of the integrated economy approach is to the role of multinational enterprise in the world economy. Multinationals fit awkwardly into perfect-competition trade theory. After all, the whole subject is concerned with the way that the boundaries of firms may cut across the boundaries of nations; yet in a perfectly competitive, constant-returns model firms are essentially invisible. As a result, attempts to introduce multinationals into trade models before 1980 often involved ad hoc assumptions, such as the idea that firms in a certain industry possessed some kind

of internationally mobile specific capital. (The fact that the expansion of multinational firms is often described as "foreign direct investment", and tracked via balance of payments statistics, may have helped confuse the picture.)

Within the integrated-economy approach, however, it is possible to offer a quite simple way of thinking about the nature and role of multinational firms.[1] This analysis [due originally to Helpman (1984b)] is, as we will see shortly, quite obviously incomplete as a story about real-world multinational enterprise; but at least it offers a reasonable first cut.

Return one last time to the pre-angel integrated economy. Let us again suppose that there are two industries, X and Y, with X consisting of many differentiated products. However, we now suppose that the X industry involves two stages of production for each variety. These might involve two stages in a physical production process (that is, the upstream stage might involve a variety-specific intermediate good, as opposed to the general-purpose intermediates discussed above); or one part of the process might involve intangible services, such as home-office activities. Without necessarily committing ourselves to the latter interpretation, let us call one activity "headquarters" and the other "production". In Figure 1.7, the vector OH represents integrated-economy inputs to headquarters, HX inputs to production of X, and XO^* inputs to Y.

In contrast to the three-industry models described before, however, we now assume that for some reason it is difficult to sell headquarters services to producers of X via arms-length transactions. The reasons may involve any of the variety of explanations invoked by industrial-organization theorists to explain vertical integration; the obvious candidate is some version of the bilateral monopoly problem, if firms must make specific investments. The details do not matter, as long as we assume that in the integrated economy each headquarters and its corresponding production unit are under common management.

Now let the angel descend one last time. He or she again divides productive factors into nations, establishing an endowment point. But there is now a further choice in the degree of punishment: can activities in different nations be under common management? If so, multinational enterprise is allowed.

Consider what happens if multinational enterprise is *not* allowed. In that case each

[1]The general idea of the model described here is that multinational enterprise arises from the combination of two things: industrial-organization motives to place two or more activities under common ownership and management, and comparative-advantage reasons to place those activities in separate countries. Obviously this story need not be told in an integrated-economy framework, although it is convenient to do so. A paper very much in the same spirit as Helpman (1984) was simultaneously and independently published by Markusen (1984); Ethier (1986) refers to the general idea that both papers embody as the Markusen–Helpman model, and I will follow his usage. Ethier's own contribution focussed on the reasons for integration of activities under common ownership – the "internalization" issue that is simply assumed by Helpman and Markusen – rather than on the reasons for geographic separation of activities.

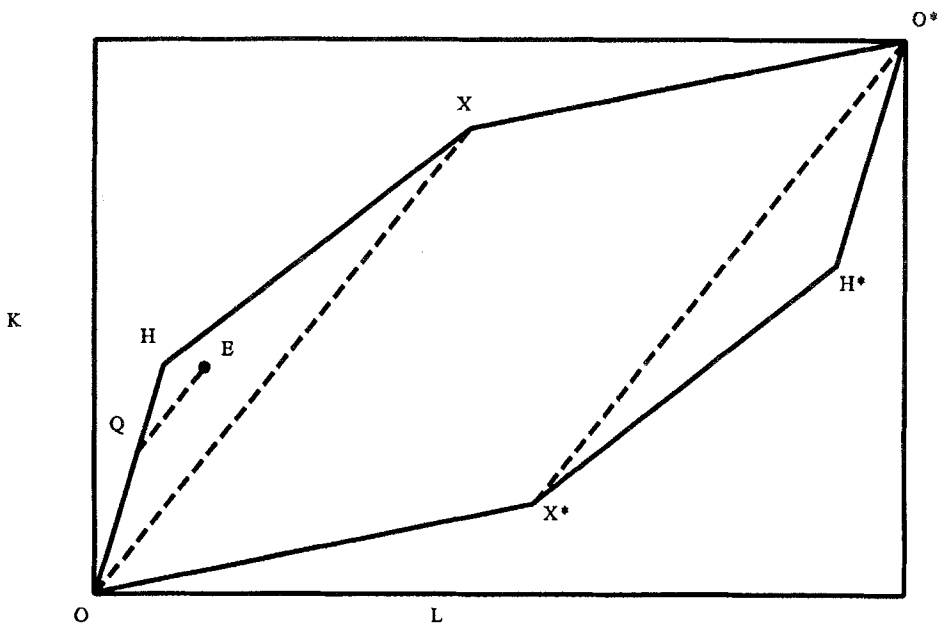

Figure 1.7.

X factory and its headquarters must be in the same country. Thus there is an X industry, whose integrated economy inputs are OX, which can be allocated between the countries (by locating some headquarters/factory combinations in each). The integrated economy can be reproduced only if the endowment point lies in the parallelogram OXO^*X^*.

If it is possible to have geographically separate operations under common control, however, the region of integrated-economy reproduction is expanded to the full hexagon $OHXO^*H^*X^*$.

Suppose that the endowment point lies in one of the "flanges" that are added by the possibility of multinational enterprise, say at point E. Then the precise structure of production – how many headquarters are located in Home – is indeterminate, although as usual the net trade in embodied factor services is determinate. It is possible to tie down the extent of multinational operation with one more assumption: that the world economy makes do with as little direct investment as possible. (This assumption may be justified by assuming that there is some small cost to operating across national boundaries, a cost small enough that the assumption of a reproduced integrated economy remains a close approximation.) Then the Home endowment will

be employed by a combination of stand-alone headquarters and integrated operations, with resources OQ devoted to the former and QE to the latter.

An important point about world equilibrium with multinationals is that trade in physical commodities will in general not be balanced. In the case shown in Figure 1.7, Home will run a persistent trade deficit. This deficit will, of course, be offset by the invisible exports of headquarters services, but one could well imagine that these exports would in practice be misreported or even ignored.

How satisfactory is this as a model of multinational enterprise? It is certainly better than trying to represent such enterprise as a movement of some kind of physical capital. Even casual observation, however, suggests that a model along these lines is very incomplete as a description of the actual extent of multinational enterprise. In the model, direct investment (by which I here mean the control of a foreign production unit by a domestic headquarters) goes in only one direction, reflecting comparative advantage. In reality, direct investment is like trade in manufactured goods: it mostly takes place among similar nations, and often reflects two-way flows within the same industry. This suggests that the integrated-economy approach is a bit less successful here than elsewhere; I will return to the issue of multinational enterprise below.

1.6. The integrated-economy approach: Concluding remarks

The parable of Samuelson's angel turns out to be a way to gain insight into a surprisingly wide range of issues in international trade: not only the original Heckscher–Ohlin–Samuelson propositions, but intraindustry trade, external economies, nontraded and intermediate goods, and even multinational enterprise. As is often the case in theory, a great simplification follows from changing the question. Instead of asking "What is the effect of unifying two economies?", we ask "What must a divided economy do to neutralize the effects of that division?". The answer, in general, is that the pieces of the integrated economy must trade factor services, explicitly or implicitly through trade in goods; and that activities which are efficient only if carried on at a single location must be concentrated in only one country. From that answer follows the result that trade reflects both comparative advantage (differences in factor endowments) and increasing returns, internal or external.

There are also two obvious limitations to the approach. One is that as a practical matter it is very clear that international transactions do *not* reproduce the integrated economy; I will return to this issue in the third part of the chapter. First, however, let us turn to the other limitation, which is that there are a number of important issues involving trade under conditions of increasing returns and imperfect competition that inherently involve market segmentation and the failure to achieve a fully integrated economy.

2. Market segmentation

2.1. The home market effect

During the 1960s and 1970s, a time when formal trade theory was almost entirely dominated by models with constant returns and perfect competition, there was a sort of "counter-culture" in international trade that claimed that other forces explained trade patterns. Perhaps the most influential work in this counter-tradition was Steffan Burenstam Linder's 1961 *An Essay on Trade and Transformation*, which argued among other things that countries tend to export goods for which they have large domestic *demand* rather than conventional supply-based comparative advantage. The mis-named "Linder hypothesis" (his family name is Burenstam Linder) seemed intuitively plausible to many observers, and generated a large if inconclusive empirical literature. The analytical underpinnings of the hypothesis, however, remained unclear. Burenstam Linder himself seems to have had in mind a process of induced innovation, in which ideas for new products are suggested by the environment. With the development of the new models of international trade, however, it became apparent that a version of the hypothesis could be given a much more mundane justification, resting on the interaction of scale economies and transport costs. [The approach described here was introduced in Krugman (1980).]

Consider a world in which labor is the only factor of production, and in which a constant share of expenditure falls on each of two industries: a constant-returns sector Y, and a Dixit–Stiglitz-type monopolistically-competitive sector X consisting of many differentiated products in which there is a constant elasticity of substitution ρ between any two varieties. Suppose also that there are two countries Home and Foreign, with labor forces L and L^*. Unlike in the integrated-economy approach, however, we suppose that there are costs of transportation between these two economies: while the constant returns good may be shipped costlessly, when a variety of the differentiated products is shipped only $\gamma < 1$ units arrive for each unit sent. (This assumption was described by Samuelson as the "iceberg" transportation model, in which goods simply melt in transit. It has two advantages as a modeling trick. First, it eliminates the need to introduce transportation as an additional sector. Second, it implies that the elasticity of demand with respect to a firm's f.o.b. price is the same as that with respect to its c.i.f. price, eliminating many potential complications.)

It turns out that in this model there is a "home market effect": the country with the larger economy tends to be a net exporter in the monopolistically-competitive sector. In particular, if L/L^* is sufficiently large all X production will be concentrated in Home.

The argument proceeds as follows. First, we note that as long as both countries produce the constant-returns good, wages will be equal. Let w be the common wage rate.

Next, let us *posit* that all production of X is concentrated in Home, and check to see whether this is in fact an equilibrium. Let μ be the share of expenditure that falls on X, and n be the number of firms that would earn zero profits if all X were produced in Home. Then each firm will have a total value of sales (including sales of goods that melt away in transit) equal to

$$S = \mu w[L + L^*]/n$$

Now consider an individual firm that begins production in Foreign. Since it will have the same marginal cost and face the same elasticity of demand as existing firms (remember the iceberg assumption), it will want to charge the same f.o.b. price as those firms. But if it charges the same f.o.b. price as the typical firm in Home, its c.i.f. price will be only γ times as high to local consumers, but $1/\gamma$ times as high to consumers abroad. But that means that the value of its sales will be[2]

$$S^* = \mu w[L\gamma^{\rho-1} + L^*\gamma^{1-\rho}]/n$$

Now the existing Home firms are, by assumption, earning zero profits. It therefore follows that this entering firm will be profitable if $S^* > S$, unprofitable if $S^* < S$.

Is concentration of all X production in Home an equilibrium? It is if, given such a concentration, an individual firm cannot enter profitably in Foreign. That is, concentration of production in Home is an equilibrium if $S^* < S$, or

$$L\gamma^{\rho-1} + L^*\gamma^{1-\rho} < L + L^*$$

On rearrangement, this yields the criterion for concentration of production,

$$\frac{L}{L^*} > \gamma^{1-\rho}$$

where the right hand side of the inequality exceeds one.

What is the intuition behind this result? Roughly, we can think of the logic as follows. Due to economies of scale, each good is produced in only one location, and sold to consumers in both. If the cost of production is the same, then the deciding factor in location is transport costs; total transportation costs are lower if production takes place in the country with the larger market.

If transport costs are high (corresponding to a low γ) or if the two countries are close in size, full concentration of the increasing-returns industry in the larger country will no longer occur. It is, however, possible to show that the larger country will still be a net exporter in this industry.

Aside from offering a relatively down-to-earth rationale for the Linder hypothesis,

[2]Bear in mind that a 1 percent increase in the price will reduce the *volume* of sales by ρ percent, but will reduce the *value* of sales by only $\rho - 1$ percent.

the analysis of the home market effect has played a significant role in recent developments in economic geography, our next topic.

2.2. The "new economic geography"

In the last few years there has been a movement toward applying concepts from the theory of international trade to the analysis of the location of population and industry *within* countries; or more precisely, within economic units characterized by a high degree of labor mobility. This is not exactly a new idea: Ohlin's 1933 book was, after all, entitled *Interregional and International Trade*, and concluded with the declaration that international trade theory is simply international location theory. In practice, however, international trade theorists have worked almost completely without reference to the ideas of location theorists and regional scientists.

This surprising lack of communication may have had much to do with the fact that regional analysis, by sheer force of empirical compulsion, has always taken the role of economies of scale both internal and external very seriously; as long as trade theory was dominated by a constant-returns tradition, the fields had little to say to each other. It is, however, surprising that trade theorists took so long to apply the tools of the new trade theory to economic geography.

During the 1980s, several urban economists, notably Fujita (1988), used the assumption of a monopolistically competitive sector producing nontraded intermediate goods to give a micro foundation to the agglomeration economies that are a necessary part of any theory of city formation. These models essentially exploited the same point made in the discussion of intermediate goods above: market-size effects on a monopolistically-competitive intermediate goods sector can have essentially the same implications as a pure external economy. These market-size externalities could then be inserted into a well-developed literature of which international economists have been oddly ignorant, that of urban systems models along the lines of Henderson (1974).

There is an older tradition in regional economics that also emphasizes market-size effects, but in a more specifically geographical context. Krugman (1991) is a simple model that attempts to capture the spirit of classic writings on economic geography such as Pred (1966) that emphasize cumulative processes of regional growth based on forward and backward linkages – but does so within a formal framework that draws on the insights and technical tricks of the new trade theory. It imagines an economy with two symmetric regions, east and west, two sectors, agriculture and manufacturing, and two factors of production, workers and farmers. Manufactures is a Dixit–Stiglitz sector, with iceberg transport costs to ship manufactured goods between regions. We now suppose, however, that workers (but not farmers) are mobile between regions, and will move to whichever region offers them a higher real wage.

The key question is whether equilibrium will involve workers equally divided

between the two regions, or will involve concentration of workers in one or the other region. It turns out that one can represent the issue in terms of a tension between one "centrifugal" force pushing toward dispersal of workers and two "centripetal" forces encouraging them to concentrate. The centrifugal force is the assumed immobility of farmers; if there were no economies of scale, workers would of course have an incentive to move toward locations where they are a relatively scarce factor. Working against this are two forces. First, whichever region has a larger population of workers will also offer a larger market for manufactured goods – and thus the home market effect will operate. Second, workers will, other things equal, receive a higher real wage if they are located close to the suppliers of manufactured goods.

These two centripetal forces can be identified with the "linkages" that played a large role in pre-1970 discussion of development economics, with the home market effect representing a backward linkage and the desire to be close to suppliers a forward linkage. And it is possible to develop a simple criterion for the sustainability of an equilibrium with concentrated manufacturing that depends in an intuitively sensible way on the levels of scale economies and transport costs, together with the share of manufacturing in expenditure.

The framework developed in that model is extremely special and unrealistic. Nonetheless, it has the useful feature of being relatively tractable in an area that has long seemed very resistant to formal modeling. And rather than displacing the traditional insights in that field, this effort at modeling seems to confirm their usefulness, even while helping clarify them. Indeed, one of the pleasant surprises of applying the methods of new trade theory to location issues is that they appear both to validate and to unify a number of seemingly disparate traditions in economic geography. The relationship to the "cumulative process" analysis of Pred and others has already been mentioned. There is another tradition in geography which attempts to explain location decisions in terms of plausible but ad hoc "market potential functions", a method created by analogy with physics, and exemplified by the work of Harris (1954). In Krugman (forthcoming) it is shown that models of economic geography based on new trade theory modeling tricks do imply the existence of market potential functions that are a little more complex than but bear a clear resemblance to those used in the geography literature.

Perhaps most interestingly, the new economic geography models appear to offer a way to integrate trade theory with two classic traditions in location theory: the land-rent theory of von Thünen and the central-place theory of Christaller (1933) and Lösch (1940). Krugman (1993) shows via simulation experiments with a dynamic, multi-region version of the model that a random initial spatial allocation of manufacturing tends to group itself into a multiple concentrations more or less evenly spaced across the landscape – a vindication of Lösch, whose central-place model had never been justified in terms of a fully specified description of individual behavior. Fujita (1994) has shown in models where all labor is mobile both how an urban center can be sustained and how a hierarchy of urban centers can emerge – a vindication of Christaller's vision of an urban system.

There have also been several papers that attempt to link the new economic geography back more directly to international trade issues per se. Venables (1993) has put forward a model in which vertical linkages between industries whose products are tradeable only at a cost create a tension between forces of specialization and of dispersion; in effect, intermediate goods play the same role that factor mobility plays in the more geographically oriented literature. Krugman and Livas (forthcoming) offer a model in which there is an interaction between trade policy and agglomeration: protectionism, by increasing the importance of domestic forward and backward linkages, raises the likelihood that a domestic agglomeration will be sustainable.

The verdict on the importance of the new economic geography is still not in. It will be obvious that I personally am very excited by it, and am inclined to messianic visions – not only of a grand union between trade and location theory, but of a linkage with current research on "self-organizing systems" in physics, chemistry, biology, and other fields. On the other hand, in fairness it should be reported that many geographers feel that the new literature is only telling them what they already knew, with a few technical gimmicks – and that the current trend among geographers proper is, if anything, away from quantitative modeling and toward a more literary and impressionistic approach.

2.3. Multinational enterprise, again

Transport costs that segment markets are also central to post-Helpman efforts to model multinational enterprise. As pointed out above, models in which firms place operations in different countries for comparative advantage reasons are unsatisfactory as a complete explanation for the actual pattern of foreign direct investment. An alternative view is that firms go multinational in order to improve their access to markets – that is, to avoid transportation costs or other barriers to trade in their products.

Brainard (1992) and Horstmann and Markusen (1992) have recently offered models of multinational enterprise in which the decision of firms to go multinational reflects a tradeoff between the loss of economies of scale associated with multiple plants and the reduction in transport costs they can achieve by producing locally for each market. These models are fairly general: factor proportions as well as transport costs may motivate overseas production, and firms can make a tradeoff between fixed and variable costs. The essence of the idea can, however, be conveyed in a simplified example.

Consider a world in which labor is the only factor of production, and in which there are two equal-sized countries. Assume that there is only one final-goods industry, and that this industry is characterized by Dixit–Stiglitz preferences, with an elasticity of substitution ρ between varieties. But now make the following additional assumptions:

(i) To produce each variety requires a fixed cost F_1 to operate a "headquarters",

and a second fixed cost F_2 to operate each "factory"; there is also a constant marginal cost c per unit of final output.

(ii) While the services of each headquarters can be shipped costlessly to any factories that a firm operates, it is costly to ship final goods. Specifically, of every unit of final good shipped internationally, only γ units arrive.

Clearly, in this setup each firm faces a choice between two strategies. It may choose to have only one factory, and export its variety to the other country; or it may incur the extra fixed cost to open a second factory, and supply each market from local production. Obviously this choice will depend on the parameters of the model.

As in the case of the home market effect, it is easiest to develop intuition about this model by positing an equilibrium of one form or the other, then checking to see whether an individual firm will have an incentive to follow the posited strategy. Suppose, then, that all firms engage in local production. Then there will in effect be a world economy in which each good is produced with a fixed labor input $F_1 + 2F_2$, and with a constant marginal labor requirement c. Firms will enter until profits are eliminated, and there will be a symmetric equilibrium with the same number of headquarters in each country; denote this number by n. Each firm will then have sales of

$$S = 2wL/n$$

with half of its sales in each country. The operating surplus of each firm – the excess of sales over the non-fixed part of its costs – will be proportional to its total sales, and equal to its fixed costs:

$$OS = S/\rho = F_1 + 2F_2$$

But is this the most profitable strategy for each individual firm? Suppose that a single firm decides to opt for producing all its output in the headquarters country. It will now face a higher marginal cost of delivering output to the other country; it is straightforward to show that its operating surplus will fall to

$$OS' = OS(1 + \gamma^{\rho-1})/2$$

But the firm will also save on fixed costs, eliminating F_2 for the second factory. Thus the firm will find it optimal to close one plant unless

$$OS - OS' = (1 - \gamma^{\rho-1})(F_1 + 2F_2)/2 > F_2$$

or

$$1 - \gamma^{\rho-1} > \frac{2}{(F_1/F_2) + 2}$$

This, then, is the criterion for sustainability of an equilibrium in which all firms are

multinational. It will tend to be satisfied, in particular, if F_2/F_1 is small, that is, if the fixed costs of opening an additional plant are not too large – or, to put it differently, if economies of scale are primarily at the level of the firm, not that of the individual plant. It will also tend to be satisfied if γ is small, that is, if transportation costs are high.

If the equilibrium does take the form of multi-plant operation, the integration of the two countries will take the following form: each country has n headquarters; each headquarters controls a plant in each country; there is no trade in final products, but there is implicit exchange of headquarter services. That is, the equilibrium will involve *intraindustry* direct foreign investment between similar countries, which is in fact characteristic of much direct investment.

Of course this model has the somewhat unsatisfactory feature that all trade is in implicit services, and that there is no trade in goods at all. It is not too hard, however, to remedy this feature by adding additional sectors and/or intermediate goods.

It may be interesting to note that the two approaches to modeling multinational enterprise described in this chapter appear to offer opposite answers to one of the traditional questions in the informal literature on direct investment: are trade and direct investment complements or substitutes? If firms go multinational in order to take advantage of cost differences – which is the underlying motive in the Markusen–Helpman model – then by so doing they will tend to create international trade. And conversely we would expect that factors that tend to increase international trade in any case, such as reductions in transportation cost, will encourage firms to separate operations geographically and thus to become multinational.

If, on the other hand, firms go multinational in order to get better access to local markets, by so doing they will replace conventional international trade. And barriers to trade, both natural and artificial, will promote such market-oriented internationalization of operations.

In the real world, there is no question that both motives are operating. Japanese electronics firms who move assembly operations to Thailand are engaging in comparative-advantage direct investment; Japanese auto firms who establish plants in the United States or the UK are engaging in market-access direct investment.

It may also be worth pointing out that in all theories of multinational enterprise it is assumed that there exist motives for placing different operations that can in principle be geographically separated under common control. The nature of these motives is, however, left fairly unspecified. The fact is that despite some deep insights by industrial organization theorists, there is still no generally accepted theory of the boundaries of the firm – certainly no theory that is operational in the sense that it can be used to predict the effects of technological or resource changes on those boundaries.

Because of these limitations, the existing models of multinational enterprise are less helpful than we would like in interpreting, or still less predicting, the effects of changes in the world economy on the extent of multinational enterprise.

2.4. Price discrimination

The models discussed in this section up to this point share many of the same features: they are general-equilibrium monopolistic-competition models, based on the two tricks of Dixit–Stiglitz preferences and iceberg transport costs. There is one more strand in the literature on trade with segmented markets, however, which needs to be discussed: that of price discrimination as a cause of trade. The models in this area are characteristically quite different, based on partial equilibrium analysis and homogeneous products.

International economists have long had to take account of the possible role of price discrimination in international trade, if only because anti-dumping legislation is of so much practical importance. Price discrimination between national markets is possible, of course, only if the markets are segmented by transport costs or other barriers. Traditional analyses of dumping, however, treated the price-discriminating firm as a pure monopolist at home, and often as a pure price-taker in export markets.

Brander (1981) first suggested modeling price discrimination in a framework in which two firms are able to sell in both of two markets, and came up with a surprising conclusion: that the firms would each sell into the other's market, possibly generating two-way trade in identical products. The point was elaborated, together with a welfare analysis, in Brander and Krugman (1983).

The basic point of the analysis may be made with a partial-equilibrium, linear example. Consider two symmetric firms producing the same good, each initially a domestic monopolist, each with the linear cost function

$$C = F + cX$$

and facing the linear demand curve

$$P = A - BX$$

Suppose initially that there is no possibility of exporting the good. Then each firm will behave as a monopolist, charging the optimal monopoly markup $(A - c)/2$.

Now suppose that the possibility of trade is opened up, but only at a cost; specifically, suppose that goods can be shipped from one market to the other only at a unit transportation cost t. If this cost is not too high – specifically, if $t < (a - c)/2$ – then each firm will have what appears to be a profitable opportunity to "dump" output in the other's market. After all, if a firm can sell a unit of its good in the other market, even after absorbing the transport cost it will still receive a price above marginal cost. It refrains from selling additional units in its own market, of course, because it is aware that to do so would drive down the price of inframarginal units; but export sales will drive down the price of someone else's inframarginal units, not its own.

But does this temptation to absorb transportation costs and sell in the other country's market persist in equilibrium? The answer depends on the nature of

competition. In the Brander analysis firms are assumed to engage in an extended version of Cournot quantity competition. Each firm chooses *two* quantities: its level of shipments to the domestic and export markets respectively. And each firm takes the other firm's shipments to those markets as given. In this case the firms can be seen as playing two separate Cournot games, one in each market.

Since these markets are symmetric, it suffices to consider one country's market. In Figure 2.1, X is the domestic firm's delivery to its own market, Y the foreign firm's exports to that market. The price in that market is determined by the demand curve

$$P = A - B(X + Y)$$

The two lines are the reaction functions of the two firms. They are derived by maximizing each firm's operating surplus in the local market, taking the other's deliveries as given. The domestic firm maximizes

$$\Pi = (P - c)X = [a - B(X + Y) - c]X$$

implying the reaction function

$$X = \frac{A - c}{2B} - \frac{Y}{2}$$

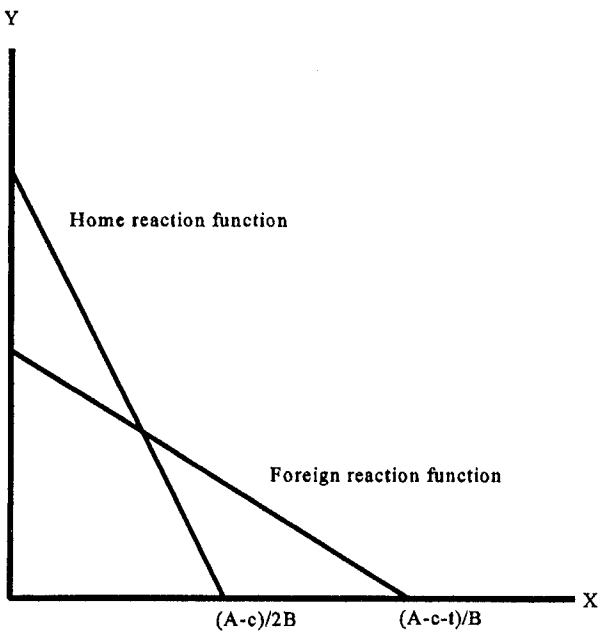

Figure 2.1.

The foreign firm, which must incur a transport cost to sell in this market, maximizes

$$\Pi^* = (P - c - t)X$$

implying the reaction function

$$Y = \frac{A - c - t}{2B} - \frac{X}{2}$$

Now recall that Y represents sales by the foreign firm in the domestic market. If these reaction functions intersect at a positive value of Y, then equilibrium will involve exports by the foreign firm to the home market – and since the markets are symmetric, exports by the home firm to the foreign market. That is, there will be two-way trade in an identical product. As the figure makes clear, this will take place as long as

$$t < \frac{A - c}{2}$$

that is, as long as the transportation cost is less than the pre-trade monopoly markup.

This model of trade due to "reciprocal dumping" is interesting in several respects. Aside from the seemingly paradoxical result of two-way trade in the same good, it is a model in which the usual roles of market structure and economies of scale are reversed. That is, in most models of trade with increasing returns and imperfect competition, the driving force behind trade is the increasing returns; imperfect competition is a necessary result of those increasing returns, but not a motivating factor. Indeed, the monopolistic-competition model is as close to perfect competition as one can get with marginal cost below average. In the reciprocal-dumping model, however, trade is essentially driven by imperfect competition – it is precisely because price is above marginal cost that each firm is tempted to raid the other's market.

How reasonable is the reciprocal dumping story? Theoretical criticism has focussed on the assumption of multimarket Cournot competition. After all, in practice oligopolistic firms seem as a rule to set prices rather than auction off predetermined quantities. The standard justification for Cournot competition in the industrial organization literature is to imagine firms that are price-setters, but are subject to short-run capacity constraints [Kreps and Scheinkman (1983)]; in a single market this ends up producing de facto quantity competition in terms of capacity choice. But even a firm that is able to price-discriminate between domestic and foreign markets is likely to face only a single, global capacity constraint. Ben-Zvi and Helpman (1992) have examined what happens to the Brander–Krugman model when the firms engage in price rather than quantity competition, with a prior competition in setting *overall* capacity. It turns out that the two-way trade in the

product disappears; the result is instead one of limit pricing, in which each firm sets a price equal to marginal cost plus transportation costs.

As an empirical matter, two-way international trade in literally identical products is surely rare. On the other hand, the practice of "basing point pricing", in which firms absorb transport costs, has been historically important in interregional trade within the United States, and has led to substantial cross-hauling of such products as cement and steel. Basing point pricing arises in the context of imperfect collusion, rather than a noncooperative game. Nonetheless, it is essentially driven by the desire of firms to raid their rivals' market areas in order to take advantage of prices above marginal cost, and is thus at least a cousin of the reciprocal dumping story about trade.

2.5. Concluding remarks

The literature on segmented markets in international trade has two main virtues. First, it helps us recognize that access to markets may matter. This point will seem hardly worth mentioning to anyone who has looked either at the pattern of actual trade flows, which fall off sharply with distance, or at surveys of businesses, who invariably mention proximity to markets as a prime determinant of location. But many models of international trade, including the integrated-economy framework applied to imperfect competition and scale economies, treat transportion and transaction costs across space as nuisances to be assumed away or sterilized by introducing a rigid distinction between traded and nontraded goods. The segmented market literature reminds us that in some cases, far from being an intellectual nuisance, the costs of doing business across space may be an important explanatory variable.

Second, the segmented-market literature offers a useful set of modeling tricks for handling the effects of such costs. In particular, the combination of Dixit–Stiglitz monopolistic competition with Samuelson's iceberg transport costs – which involves a kind of layering of implausible but convenient assumptions – has turned out to be a remarkably flexible tool of analysis, allowing us to think coherently about market-size and market-access effects in a wide variety of contexts.

3. Unresolved issues and future concerns

This final section of the chapter considers a number of issues that the literature on increasing returns and imperfect competition in trade has either not addressed effectively or not addressed at all, but that seem of considerable importance. I consider these issues under two headings: issues regarding the pattern of specialization and trade, and those regarding more fundamental micro foundations.

3.1. Trade issues per se

3.1.1. Failure to reproduce the integrated economy

Focussing on trade that reproduces an integrated economy turns out to be a tremendously powerful simplifying and unifying device, and did much to make the literature on increasing returns and imperfect competition seem like a natural next step in the development of trade theory rather than a set of disparate alternative approaches. Unfortunately, casual observation suggests pretty strongly that trade does *not* reproduce the integrated economy. Wage rates of equivalent labor are certainly not equalized.

While transportation costs are one reason for a breakdown in factor-price equalization, few economists think that they are the main explanation of this incomplete integration. Rather, most of us are inclined to blame one or both of two causes: differences in factor endowments which are too large to allow factor price equalization, and differences in production functions between countries. These may be summarized by saying that the world economy is either out of the parallelogram, or completely outside the box.

There is no fundamental conceptual difficulty in introducing either wide endowment disparities or technological differences into models with increasing returns and either monopolistic competition or external economies. What is lacking, however, is any systematic treatment of the patterns of specialization that might result, or of the likely empirical counterparts of such deviations from the canonical assumptions. This criticism admittedly applies equally to constant-returns trade theory; indeed, many international economists seem remarkably relaxed about relying on Heckscher–Ohlin models to analyze some issues, Ricardian models to analyze others (in particular, the trade foundations of international macroeconomic models are generally more or less Ricardian).

There has been some movement toward modeling that combines factor proportions with technological differences. Davis (1992) has offered an influential critique of the interpretation of intraindustry trade as scale-economy-based, proposing instead that it is due to exogenous technological differences. But it would be useful if there were a more systematic treatment of what happens when the integrated-economy approach doesn't work.

3.1.2. Multilateral trade

The great bulk of the theoretical literature on international trade, old and new, focusses on trade between two countries. In the real world, however, there are many countries, and practical policy debates in international economics often depend on the interpretation of multilateral trade data. Two important recent examples are the dispute over the degree of closure of Japan's market, in which many authors tried to

compare Japan's actual volume of trade with the predicted volume given an empirical model of multilateral trade [see, for example, Lawrence (1987)]; and the more recent literature in which essentially the same method is used to ask whether East Asia is becoming an implicit trading bloc (see Frankel 1993).

In both literatures the essential empirical tool was some version of the "gravity" equation, which relates the trade between any pair of countries to their populations, incomes, and the distance between them. But some trade theorists have complained that such analysis is irrelevant, because the gravity equations do not arise from any well-specified model of multilateral trade.

At first sight, monopolistic competition models of trade seem to offer a justification for using gravity-type regressions. Helpman and Krugman (1985) consider the case of a world in which *all* tradeable industries consist of differentiated products. If tastes are identical and homothetic, and – crucially – if those goods that can be traded have zero transport costs, then the value of exports from any country j to some other country k will be

$$T_{jk} = \tau \frac{Y_j Y_k}{Y_W}$$

where Y_j is the GDP of country j, Y_W is gross world product, and τ is the share of tradeables in expenditure. This looks like a classic gravity equation: trade between two countries, other things equal, is proportional to the product of the incomes of those two countries. And this analysis helps to suggest why gravity equations work as well as they do. Unfortunately, empirical gravity equations invariably include not only the incomes of j and k but some measure of the distance between the countries – and distance always shows up as a crucial determinant of trade flows. That is, the empirical evidence, reflected in the empirical models, clearly shows that transportation costs (perhaps in the form of invisible transaction costs) play a crucial role in the pattern of multilateral trade.[3]

The problem, however, is that the apparent rationale for a simple gravity equation breaks down when transport costs are important. Even in a one-industry world, the trade between two countries should depend not only on their incomes and the distance between them but on the sizes and distances of other economies. Consider, for example, two small equal-size economies. If they were located on Mars, able to trade with each other but separated from all other economies by near-prohibitive transport costs, half of each country's spending on tradeables should consist of imports from

[3]An important practical issue is what we mean by "transport costs". Measured shipping costs are quite small for most goods that can be shipped at all; yet trade falls off quite strongly with distance. This suggests that transportation costs in our models are a proxy for more subtle transaction costs involving the difficulty of maintaining personal contact, or perhaps differences in culture that are correlated with physical distance. In any case, one wonders how badly misleading it is to represent these as a proportional melting of goods in transit!

the other. That is, the value of trade between the two should be equal to $\tau Y/2$, where Y is each country's income. On the other hand, if the two countries were located in the middle of Europe, close to much larger countries offering competing products, then each would spend only a small fraction of its income on imports from the other. Thus the trade between them would be much less, even if the distance between the two were the same.

Matters become much worse when we consider that there are many sectors, differing both in the importance of scale economies and in the level of transport cost. Now the pattern of trade may be influenced by home market effects; sectors characterized by large economies of scale, for example, will tend to be produced not only by countries that are themselves large (Germany) but by those that are located close to large external markets (Belgium) – and predicting the pattern and volume of trade becomes yet more complex.

Or so it appears. Perhaps, like so many issues in economic theory, this issue is much less complex than it seems when looked at the right way. The point is that there is not now a careful, generally accepted analysis of the pattern of multilateral trade when transport costs matter – as they clearly do – which allows us to assess observed patterns with any assurance. This is not a critique of the gravity-equation based literature. The fact is that such equations work very well, and it is entirely reasonable to use an ad hoc approach that seems to work until someone comes along with a more rigorous approach. What is clear is that the lack of a good analysis of multilateral trade in the presence of transport costs is a major gap in trade theory.

3.2. Deeper issues

Ohlin concluded his 1933 book by declaring that international trade theory is nothing but international location theory. The new economic geography is in a way an attempt to prove his point, but he was surely only partly right: if there is an overriding conclusion from the last 15 years of research it is that international trade theory is also international industrial organization. And therefore any weaknesses in our understanding of industrial organization are also weaknesses in our study of international economics.

Where is our theory of industrial organization most inadequate? The answer, surely, is in the attempt to explain the nature and boundaries of the firm. Why are some activities carried out within hierarchical organizations, while others are carried out through arms-length transactions? There is a rigorous literature demonstrating that the simple view that vertical integration is a response to market imperfections is not a sufficient answer, because it does not explain why individuals do not write contracts to deal with these imperfections. This result is, however, essentially negative; efforts to explain, given this result, why some transactions characteristically take place via markets and others within command structures are at best suggestive. Indeed, there is

even a dispute over whether the conventional distinction between transactions carried out within and between firms is truly valid. Some theorists, such as Michael Jensen, claim that individuals are the only true economic actors, and that the economy is simply a web of contracts; calling some dense parts of that web firms is no more than a conceptual and legal convenience. On the other hand, such theorists as Oliver Hart have argued that the distinction between within-firm and market transactions is indeed a fundamental one, and that ownership of productive assets is not just a particular form of contract but a crucial issue.

What is noteworthy, however, is that the theory of trade under conditions of increasing returns and imperfect competition is at present blithely insensitive to these conceptual problems. Consider even the simplest monopolistic-competition model of trade. Why do we assume that each firm produces only a single variety, rather than imagine a firm or alliance of firms producing many varieties and taking their interdependence into account in pricing decisions? Or for that matter why don't firms use two-part tariffs to allow customers to pay marginal cost? The answer is that we have implicit notions that there are diseconomies to grouping many products into a single firm, and that the transactions costs of nonlinear pricing schemes are too high. In this case these implicit assumptions seem reasonable, but the models are actually less rigorous than they first appear.

More seriously, when there are external economies, is it right to assume that firms behave atomistically – or should we assume that someone will try to internalize the externality? (Interestingly and contrastingly, in Henderson (1974)-type urban models the standard closure is to envision "city corporations" that compete precisely by internalizing the external economies of agglomeration.) Again, the assumption that firms remain small and take external effects as given represents some implicit theorizing.

Where the failure to have a real theory of the boundaries of the firm becomes truly serious, however, is of course in the analysis of multinational firms. While Ethier and others have tried to establish microfoundations for the internalization decision, these efforts are, like those in the industrial organization literature generally, interesting and suggestive rather than fully satisfactory. Why, exactly, did United Fruit want to own Central American banana plantations (and often the republics in which they were located), while many US sellers of personal computer clones seem reconciled simply to contract with their Korean or Taiwanese suppliers? The answer is not at all obvious from the international economics literature.

What makes this an important issue is, in particular, the fact that firms in today's international economy are in the process of experimenting both with new boundaries and with novel forms of organization. During the late 1980s Japanese electronics manufacturers managed, with a little help from fast-talking Americans, to persuade themselves that they needed to own major Hollywood film studios. Were they, despite current appearances, right? More to the point, what can our theory say about the issue? At the same time, there is a great deal of buzz in the business strategy area

about international corporate alliances – a form of interaction that is neither arms'-length market transactions nor extension of the boundaries of the firm in the usual sense. How does this fit into our analysis?

These are deep questions, and not specifically international in nature. Nonetheless, international as well as general industrial organization must eventually confront them.

4. Concluding remarks

I do not want to end this survey on a negative note. The previous section was, in effect, a catalogue of problems with what has become the standard analysis of trade under increasing returns and imperfect competition. In spite of these problems, however, the most remarkable thing to someone who was trained in international trade theory before 1980 – or even to someone who read the 1982 Helpman survey – is how comprehensible the roles of increasing returns and imperfect competition have turned out to be. Before the new models appeared, traditional trade theorists believed that to introduce these factors into the theory would be to plunge into an impenetrable thicket of conceptual difficulties. When it turned out that coherent models could be written down, the expectation was that they would be too diffuse a set of mutually contradictory approaches to offer more than isolated insights. Instead, the field has turned out to be characterized by a surprising degree both of cohesion and of continuity with the older traditions of international trade theory.

References

Ben-Zvi, S. and E. Helpman (1992), "Oligopoly in segmented markets", in: G. Grossman, ed., Imperfect competition and international trade (MIT Press, Cambridge, MA) 31–53.

Brainard, L. (1992), "A simple theory of multinational corporations and trade with a trade-off between proximity and concentration", mimeo (MIT Press, Cambridge, MA).

Brander, J. (1981), "Intra-industry trade in identical commodities", Journal of International Economics 11:1–14.

Brander, J. and P. Krugman (1983), "A 'reciprocal dumping' model of international trade", Journal of International Economics 15:313–323.

Burenstam Linder, S. (1961), An essay on trade and transformation (Wiley, New York).

Christaller, W. (1933), Central places in southern Germany [English translation by C.W. Baskin (1966) (Prentice-Hall, Englewood Cliffs)].

Davis, D. (1992), "Intra-industry trade: A Heckscher–Ohlin–Ricardo approach", in: Essays in the theory of international trade and economic growth, Ph.D. thesis, Columbia University.

Dixit, A. and V. Norman (1980), Theory of international trade (Cambridge University Press, Cambridge, UK).

Dixit, A. and J. Stiglitz (1977), "Monopolistic competition and optimum product diversity", American Economic Review 67:297–308.

Ethier, W. (1979), "Internationally decreasing costs and world trade", Journal of International Economics 9:1–24.

Ethier, W. (1982a), "Decreasing costs in international trade and Frank Graham's argument for protection", Econometrica 50:1243–1268.
Ethier, W. (1982b), "National and international returns to scale in the modern theory of international trade", American Economic Review 72:950–959.
Ethier, W. (1986), "The multinational firm", Quarterly Journal of Economics 101:805–834.
Frankel, J. (1993), "Is Japan creating a yen bloc in East Asia and the Pacific?", in: J. Frankel and M. Kahler, eds., Regionalism and rivalry: Japan and the U.S. in Pacific Asia (University of Chicago Press, Chicago).
Fujita, M. (1988), Urban economic theory (Oxford University Press, Oxford).
Fujita, M. (1994), Monopolistic competition and urban hierarchies, mimeo, University of Pennsylvania.
Harris, C.D. (1954), "The market as a factor in the localization of industry in the U.S.", Annals of the Association of American Geographers 44:315–348.
Helpman, E. (1981), "International trade in the presence of product differentiation, economies of scale, and imperfect competition: A Chamberlin–Heckscher–Ohlin approach", Journal of International Economics 11:305–340.
Helpman, E. (1984a), "Increasing returns, imperfect markets, and trade theory", in: R. Jones and P. Kenen, eds., Handbook of international economics, vol. 1 (North-Holland, Amsterdam).
Helpman, E. (1984b), "A simple theory of trade with multinational corporations", Journal of Political Economy 92:451–472.
Helpman, E. and P. Krugman (1985), Market structure and foreign trade (MIT Press, Cambridge, MA).
Henderson, J.V. (1974), "On the sizes and types of cities", American Economic Review 64:640–656.
Horstmann, I. and J. Markusen (1992), "Endogenous market structures in international trade", Journal of International Economics 32:109–129.
Kreps, D. and J. Scheinkman (1983), "Quantity precommitment and Bertrand competition yield Cournot outcomes", Bell Journal of Economics 12:326–337.
Krugman, P. (1980), "Scale economies, product differentiation, and the pattern of trade", American Economic Review 70:950–959.
Krugman, P. (1991), "Increasing returns and economic geography", Journal of Political Economy 99:183–199.
Krugman, P. (1993), "On the number and location of cities", European Economic Review 37:293–298.
Krugman, P. (1995) Development, geography, and economic theory (MIT Press, Cambridge, MA), forthcoming.
Krugman, P. and R. Livas-Elizondo (1995), "Trade policy and the third world metropolis", Journal of Political Economy, forthcoming.
Lancaster, K. (1975), "Socially optimal product differentiation", American Economic Review 65:567–585.
Lawrence, R. (1987), "Japan: Closed markets or minds?", Brookings Papers on Economic Activity 2:517–554.
Lösch, A. (1940), The economics of location [English translation (1954) (Yale University Press, New Haven)].
Markusen, J. (1984), "Multinationals, multi-plant economies, and the gains from trade", Journal of International Economics 16:205–226.
Pred, A. (1966), The spatial dynamics of U.S. urban-industrial growth, 1800–1914 (MIT Press, Cambridge, MA).
Romer, P. (1986), "Increasing returns and long-run growth", Journal of Political Economy 94:1002–1037.
Venables, A. (1993), "Equilibrium locations of vertically linked industries", mimeo, London School of Economics.