Journal of International Economics 9 (1979) 469-479. © North-Holland Publishing Company

INCREASING RETURNS, MONOPOLISTIC COMPETITION, AND INTERNATIONAL TRADE

Paul R. KRUGMAN

Yale University, New Haven, CT 06520, USA

Received November 1978, revised version received February 1979

This paper develops a simple, general equilibrium model of noncomparative advantage trade. Trade is driven by economies of scale, which are internal to firms. Because of the scale economies, markets are imperfectly competitive. Nonetheless, one can show that trade, and gains from trade, will occur, even between countries with identical tastes, technology, and factor endowments.

1. Introduction

It has been widely recognized that economies of scale provide an alternative to differences in technology or factor endowments as an explanation of international specialization and trade. The role of 'economies of large scale production' is a major subtheme in the work of Ohlin (1933); while some authors, especially Balassa (1967) and Kravis (1971), have argued that scale economies play a crucial role in explaining the postwar growth in trade among the industrial countries. Nonetheless, increasing returns as a cause of trade has received relatively little attention from formal trade theory. The main reason for this neglect seems to be that it has appeared difficult to deal with the implications of increasing returns for market structure.

This paper develops a simple formal model in which trade is caused by economies of scale instead of differences in factor endowments or technology. The approach differs from that of most other formal treatments of trade under increasing returns, which assume that scale economies are external to firms, so that markets remain perfectly competitive.[1] Instead, scale economies are here assumed to be internal to firms, with the market structure that emerges being one of Chamberlinian monopolistic competition.[2] The formal

[1]Authors who allow for increasing returns in trade by assuming that scale economies are external to firms include Chacoliades (1970), Melvin (1969), and Kemp (1964), and Negishi (1969).

[2]A Chamberlinian approach to international trade is suggested by Gray (1973). Negishi (1972) develops a full general-equilibrium model of scale economies, monopolistic competition, and trade which is similar in spirit to this paper, though far more complex. Scale economies and product differentiation are also suggested as causes of trade by Barker (1977) and Grubel (1970).

treatment of monopolistic competition is borrowed with slight modifications from recent work by Dixit and Stiglitz (1977). A Chamberlinian formulation of the problem turns out to have several advantages. First, it yields a very simple model; the analysis of increasing returns and trade is hardly more complicated than the two-good Ricardian model. Secondly, the model is free from the multiple equilibria which are the rule when scale economies are external to firms, and which can detract from the main point. Finally, the model's picture of trade in a large number of differentiated products fits in well with the empirical literature on 'intra-industry' trade [e.g. Grubel and Lloyd (1975)].

The paper is organized as follows. Section 2 develops the basic modified Dixit-Stiglitz model of monopolistic competition for a closed economy. Section 3 then examines the effects of opening trade as well as the essentially equivalent effects of population growth and factor mobility. Finally, section 4 summarizes the results and suggests some conclusions.

2. Monopolistic competition in a closed economy

This section develops the basic model of monopolistic competition with which I will work in the next sections. The model is a simplified version of the model developed by Dixit and Stiglitz. Instead of trying to develop a general model, this paper will assume particular forms for utility and cost functions. The functional forms chosen give the model a simplified structure which makes the analysis easier.

Consider, then, an economy with only one scarce factor of production, labor. The economy is assumed able to produce any of a large number of goods, with the goods indexed by i. We order the goods so that those actually produced range from 1 to n, where n is also assumed to be a large number, although small relative to the number of potential products.

All residents are assumed to share the same utility function, into which all goods enter symmetrically.

$$U = \sum_{i=1}^{n} u(c_i), \qquad u' > 0, \quad u'' < 0, \tag{1}$$

where c_i is the consumption of the ith good.

It will be useful to define a variable, ε, where

$$\varepsilon_i = -\frac{u'}{u'' c_i}. \tag{2}$$

and where we assume $\partial \varepsilon_i / \partial c_i < 0$. The variable ε_i will turn out to be the

elasticity of demand facing an individual producer; the reasons for assuming that is is decreasing in c_i will become apparent later.

All goods are also assumed to be produced with the same cost function. The labor used in producing each good is a linear function of output,

$$l_i = \alpha + \beta x_i, \qquad \alpha, \beta > 0, \tag{3}$$

where l_i is labor used in producing good i, x_i is the output of good i, and α is a fixed cost. In other words, there are decreasing average costs and constant marginal costs.

Production of a good must equal the sum of individual consumptions of the good. If we identify individuals with workers, production must equal the consumption of a representative individual times the labor force:

$$x_i = Lc_i. \tag{4}$$

Finally, we assume full employment, so that the total labor force L must be exhausted by employment in production of individual goods:

$$L = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} [\alpha + \beta x_i]. \tag{5}$$

Now there are three variables we want to determine: the price of each good relative to wages, p_i/w; the output of each good, x_i; and the number of goods produced, n. The symmetry of the problem will ensure that all goods actually produced will be produced in the same quantity and at the same price, so that we can use the shorthand notation

$$\left.\begin{aligned} p &= p_i \\ x &= x_i \end{aligned}\right\}, \quad \text{for all } i. \tag{6}$$

We can proceed in three stages. First, we analyze the demand curve facing an individual firm; then we derive the pricing policy of firms and relate profitability to output; finally, we use an analysis of profitability and entry to determine the number of firms.

To analyze the demand curve facing the firm producing some particular product, consider the behavior of a representative individual. He will maximize his utility (1) subject to a budget constraint. The first-order conditions from that maximization problem have the form

$$v'(c_i) = \lambda p_i, \qquad i = 1, \ldots, n, \tag{7}$$

where λ is the shadow price on the budget constraint, which can be interpreted as the marginal utility of income.

We can substitute the relationship between individual consumption and output into (7) to turn it into an expression for the demand facing an individual firm,

$$p_i = \lambda^{-1} v'(x_i/L). \tag{8}$$

If the number of goods produced is large, each firm's pricing policy will have a negligible effect on the marginal utility of income, so that it can take λ as fixed. In that case the elasticity of demand facing the ith firm will, as already noted, be $\varepsilon_i = -v'/v''c_i$.

Now let us consider profit-maximizing pricing behavior. Each individual firm, being small relative to the economy, can ignore the effects of its decisions on the decisions of other firms. Thus, the ith firm will choose its price to maximize its profits,

$$\Pi_i = p_i x_i - (\alpha + \beta x_i)w. \tag{9}$$

The profit-maximizing price will depend on marginal cost and on the elasticity of demand:

$$p_i = \frac{\varepsilon}{\varepsilon - 1}\beta w \tag{10}$$

or $p_i/w = \beta\varepsilon/(\varepsilon - 1)$.

Now this does not determine the price, since the elasticity of demand depends on output; thus, to find the profit-maximizing price we would have to derive profit-maximizing output as well. It will be easier, however, to determine output and prices by combining (10) with the condition that profits be zero in equilibrium.

Profits will be driven to zero by entry of new firms. The process is illustrated in fig. 1. The horizontal axis measures output of a representative firm; the vertical axis revenue and cost expressed in wage units. Total cost is shown by TC, while OR and OR^1 represent revenue functions. Suppose that given the initial number of firms, the revenue function facing each firm is given by OR. The firm will then choose its output so as to set marginal revenue equal to marginal cost, at A. At that point, since price (average revenue) exceeds average cost, firms will make profits. But this will lead entrepreneurs to start new firms. As they do so, the marginal utility of income will rise, and the revenue function will shrink in. Eventually equilibrium will be reached at a point such as B, where it is true both that marginal revenue equals marginal cost and that average revenue equals

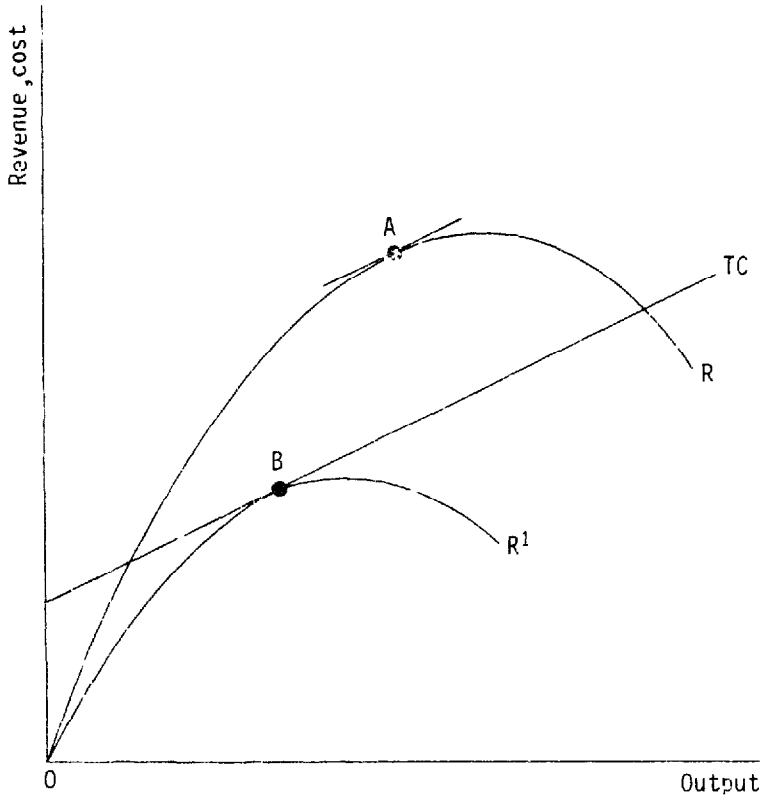

Fig. 1.

average cost. This is, of course, Chamberlin's famous tangency solution [Chamberlin (1962)].

To characterize this equilibrium more carefully, we need to show how the price and output of a representative firm can be derived from cost and utility functions. In fig. 2 the horizontal axis shows *per-capita* consumption of a representative good, while the vertical axis shows the price of a representative good in wage units. We have one relationship between c and p/w in the pricing condition (10), which is shown as the curve PP. Price lies everywhere above marginal cost, and increases with c because, by assumption, the elasticity of demand falls with c.

A second relationship between p/w and c can be derived from the condition of zero profits in equilibrium. From (9), we have

$$0 = px - (\alpha + \beta x)w, \tag{11}$$

which can be rewritten

$$p/w = \beta + \alpha/x = \beta + \alpha/Lc. \tag{12}$$

This is a rectangular hyperbola above the line $p/w = \beta$, and is shown in fig. 2 as ZZ.

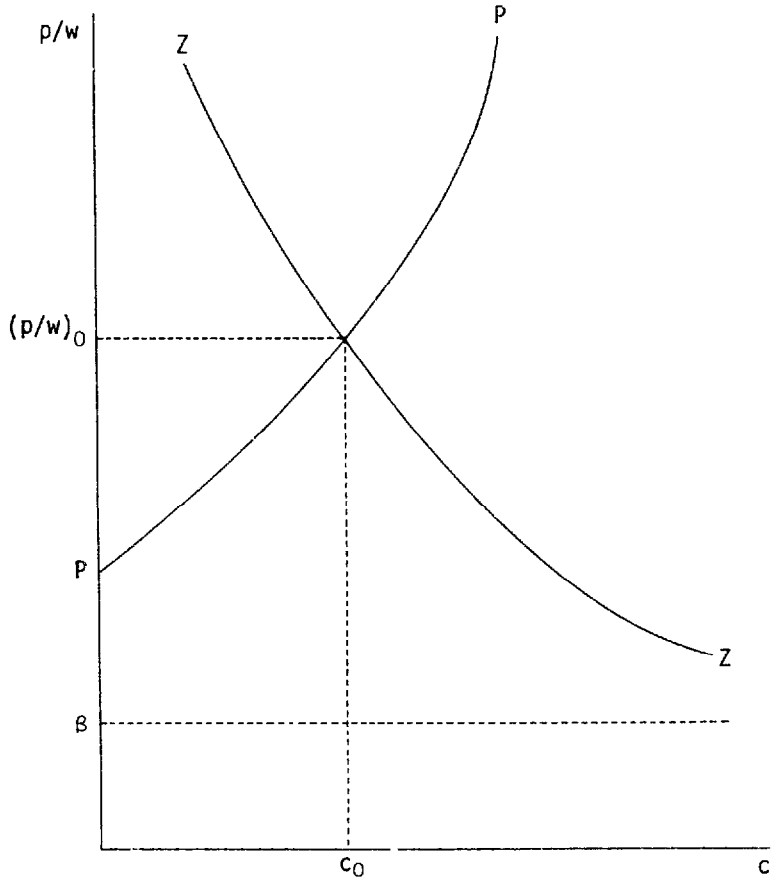

Fig. 2.

The intersection of the PP and ZZ schedules determines individual consumption of each good and the price of each good. From the consumption of each good we have output per firm, since $x = Lc$. And the assumption of full employment lets us determine the number of goods produced:

$$n = \frac{L}{\alpha + \beta x}. \tag{13}$$

We now have a complete description of equilibrium in the economy. It is indeterminate *which* n goods are produced, but it is also unimportant, since the goods enter into utility and cost symmetrically. We can now use the model to analyze the related questions of the effects of growth, trade, and factor mobility.

3. Growth, trade, and factor mobility

The model developed in the last section was a one-factor model, but one

in which there were economies of scale in the use of that factor, so that in a real sense the division of labor was limited by the extent of the market. In this section we consider three ways in which the extent of the market might increase: growth in the labor force, trade, and migration.

3.1. *Effects of labor force growth*

Suppose that an economy of the kind analyzed in the last section were to experience an increase in its labor force. What effect would this have? We can analyze some of the effects by examining fig. 3. The PP and ZZ

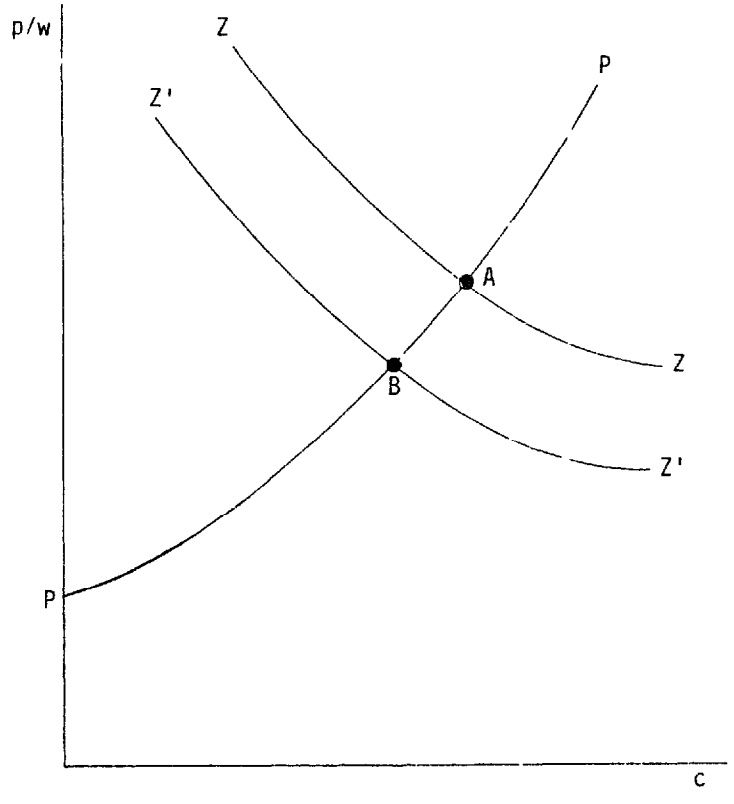

Fig. 3.

schedules have the same definitions as in fig. 2; before the increase in the labor force equilibrium is at A. By referring back to eqs. (10) and (11) we can see that an increase in L has no effect on PP, but that it causes ZZ to shift left. The new equilibrium is at B: c falls, and so does p/w. We can show, however, that both the output of each good and the number of goods produced rise. By rearranging (12) we have

$$x = \alpha/(p/w - \beta), \tag{14}$$

which shows that output must rise, while since $n = L/(\alpha + \beta Lc)$, a rise in L and a fall in c imply a rise in n.

Notice that these results depend on the fact that the *PP* curve slopes upward, which in turn depends on the assumption that the elasticity of demand falls with c. This assumption, which might alternatively be stated as an assumption that the elasticity of demand rises when the price of a good is increased, seems plausible. In any case, it seems to be necessary if this model is to yield reasonable results, and I make the assumption without apology.

We can also consider the welfare implications of growth. Comparisons of overall welfare would be illegitimate, but we can look at the welfare of representative individuals. This rises for two reasons: there is a rise in the 'real wage' w/p, and there is also a gain from increased choice, as the number of available products increases.

I have considered the case of growth at some length, even though our principal concern is with trade, because the results of the analysis of growth will be useful next, when we turn to the analysis of trade.

3.2. *Effects of trade*

Suppose there exist two economies of the kind analyzed in section 2, and that they are initially unable to trade. To make the point most strongly, assume that the countries have identical tastes and technologies. (Since this is a one-factor model, we have already ruled out differences in factor endowments.) In a conventional model, there would be no reason for trade to occur between these economies, and no potential gains from trade. In this model, however, there will be both trade and gains from trade.

To see this, suppose that trade is opened between these two economies at zero transportation cost. Symmetry will ensure that wage rates in the two countries will be equal, and that the price of any good produced in either country will be the same. The effect will be the same as if *each* country had experienced an increase in its labor force. As in the case of growth in a closed economy, there will be an increase both in the scale of production and in the range of goods available for consumption. Welfare in both countries will increase, both because of higher w/p and because of increased choice.

The direction of trade – which country exports which goods – is indeterminate; all that we can say is that each good will be produced only in one country, because there is (in this model) no reason for firms to compete for markets. The *volume* of trade, however, is determinate. Each individual will be maximizing his utility function, which may be written

$$U = \sum_{i=1}^{n} v(c_i) + \sum_{i=n+1}^{n+n^*} v(c_i), \tag{15}$$

where goods $1,\ldots,n$ are produced in the home country and $n+1,\ldots,n+n^*$ in the foreign country. The number of goods produced in each country will be proportional to the labor forces:

$$n=\frac{L}{\alpha+\beta x},$$
$$n^*=\frac{L^*}{\alpha+\beta x}. \tag{16}$$

Since all goods will have the same price, expenditures on each country's goods will be proportional to the country's labor force. The share of imports in home country expenditures, for instance, will be $L^*/(L+L^*)$; the values of imports of each country will be national income times the import share, i.e.

$$\begin{aligned} M &= wL\cdot L^*/(L+L^*) \\ &= wLL^*/(L+L^*) \\ &= M^*. \end{aligned} \tag{17}$$

Trade is balanced, as it must be, since each individual agent's budget constraint is satisfied. The volume of trade as a fraction of world income is maximized when the economies are of equal size.

We might note that the result that the volume of trade is determinate but the direction of trade is not is very similar to the well-known argument of Linder (1961). This suggests an affinity between this model and Linder's views, although Linder does not explicitly mention economies of scale.

The important point to be gained from this analysis is that economies of scale can be shown to give rise to trade and to gains from trade even when there are no international differences in tastes, technology, or factor endowments.

3.3. *Effects of factor mobility*[3]

An interesting extension of the model results when we allow for movement of labor between countries or regions. There is a parallel here with Heckscher-Ohlin theory. Mundell (1957) has shown that in a Heckscher Ohlin world trade and factor mobility would be substitutes for one another.

[3]The results in this section bear some resemblance to some nontheoretical accounts of the emergence of backward regions. We might propose the following modification of the model suppose that the population of each region is divided into a mobile group and an immobile group. Migration would then move all the mobile people to one region, leaving behind an immiserized 'Appalachia' of immobile people whose standard of living is depressed by the smallness of the market.

and that factor movements would be induced by impediments to trade such as tariffs or transportation costs. The same kinds of results emerge from this model.

To see this, suppose that there are two regions of the kind we have been discussing, and that they have the same tastes and technologies. There is room for mutual gains from trade, because the combined market would allow both greater variety of goods and a greater scale of production. The same gains could be obtained without trade, however, if the population of one region were to migrate to the other. In this model, trade and growth in the labor force are essentially equivalent. If there are impediments to trade, there will be an incentive for workers to move to the region which already has the larger labor force. This is clearest if we consider the extreme case where no trade in goods is possible, but labor is perfectly mobile. Then the more populous region will offer both a greater real wage w/p and a greater variety of goods, inducing immigration. In equilibrium all workers will have concentrated in one region or the other. Which region ends up with the population depends on initial conditions; in the presence of increasing returns history matters.

Before proceeding further we should ask what aspect of reality, if any, is captured by the story we have just told. In the presence of increasing returns factor mobility appears to produce a process of agglomeration. If we had considered a many-region model the population would still have tended to accumulate in only one region, which we may as well label a city; for this analysis seems to make most sense as an account of the growth of metropolitan areas. The theory of urban growth suggested by this model is of the 'city lights' variety: people migrate to the city in part because of the greater variety of consumption goods it offers.

Let us return now to the two-region case to make a final point. We have seen that which region ends up with the population depends on the initial distribution of population. As long as labor productivity is the same in both regions, though, there is no difference in welfare between the two possible outcomes. If there is any difference in the conditions of production between the two regions, however, it does matter which gets the population – and the process of migration can lead to the wrong outcome.

Consider, for example, a case in which both fixed and variable labor costs are higher in one region. Then it is clearly desirable that all labor should move to the other region. But if the inferior region starts with a large enough share of the population, migration may move in the wrong direction.

To summarize: in the model of this paper, as in some more conventional trade models, factor mobility can substitute for trade. If there are impediments to trade, labor will concentrate in a single region; which region depends on the initial distribution of population. Finally, the process of agglomeration may lead population to concentrate in the wrong place.

4. Summary and conclusions

This paper adapts a Chamberlinian approach to the analysis of trade under conditions of increasing returns to scale. It shows that trade need not be a result of international differences in technology or factor endowments. Instead, trade may simply be a way of extending the market and allowing exploitation of scale economies, with the effects of trade being similar to those of labor force growth and regional agglomeration. This is a view of trade which appears to be useful in understanding trade among the industrial countries.

What is surprising about this analysis is that it is extremely simple. While the role of economies of scale in causing trade has been known for some time, it has been underemphasized in formal trade theory (and in textbooks). This paper shows that a clear, rigorous, and one hopes persuasive model of trade under conditions of increasing returns can be constructed. Perhaps this will help give economies of scale a more prominent place in trade theory.

References

Balassa, Bela, 1967, Trade liberalization among industrial countries (McGraw-Hill, New York).

Barker, Terry, 1977, International trade and economic growth: An alternative to the neoclassical approach, Cambridge Journal of Economics 1, no. 2, 153–172.

Chacoliades, Miltiades, 1970, Increasing returns and the theory of comparative advantage, Southern Economic Journal 37, no. 2, 157–162.

Chamberlin, Edward, 1962, The theory of monopolistic competition.

Dixit, Avinash and Joseph Stiglitz, 1977, Monopolistic competition and optimum product diversity, American Economic Review, June, 297–308.

Gray, Peter, 1973, Two-way international trade in manufactures: A theoretical underpinning, Weltwirtschaftliches Archiv 109, 19–39.

Grubel, Herbert, 1970, The theory of intra-industry trade, in: I.A. McDougall and R.H. Snape, eds., Studies in international economics (North-Holland, Amsterdam).

Grubel, Herbert and Peter Lloyd, 1975, Intra-industry trade (MacMillan, London).

Hufbauer, Gary and John Chilas, 1974, Specialization by industrial countries: Extent and consequences, in: H. Giersch, ed., The international division of labour (Institut für Weltwirtschaft, Kiel).

Kemp, Murray, 1964, The pure theory of international trade (Prentice-Hall).

Kindleberger, Charles, 1973, International economics (Irwin).

Kravis, Irving, 1971, The current case for import limitations, in: Commission on International Trade and Investment Policy, United States Economic Policy in an Interdependent World (U.S. Government Printing Office, Washington).

Linder, Staffan Burenstam, 1961, An essay on trade and transformation (John Wiley and Sons).

Melvin, James, 1969, Increasing returns to scale as a determinant of trade, Canadian Journal of Economics and Political Science 2, no. 3, 389–402.

Mundell, Robert, 1957, International trade and factor mobility, American Economic Review 47, 321–335.

Negishi, Takashi, 1969, Marshallian external economies and gains from trade between similar countries, Review of Economic Studies 36, 131–135.

Negishi, Takashi, 1972, General equilibrium theory and international trade (North-Holland, Amsterdam).

Ohlin, Bertil, 1933, Interregional and international trade (Harvard University Press).