Optimal Integration Strategies for the Multinational Firm*

by

Gene M. Grossman
Princeton University

Elhanan Helpman
Harvard University, Tel Aviv University, and CIAR

and

Adam Szeidl
Harvard University

April 27, 2004

Abstract

We examine integration strategies of multinational firms that face a rich array of choices of international organization. Each firm in an industry must provide headquarter services from its home country, but can produce its intermediate inputs and conduct assembly operations in one or more of three locations. We study the equilibrium choices of firms that differ in productivity levels, focusing on the role that industry characteristics such as the fixed costs of foreign subsidiaries, the cost of transporting intermediate and final goods, and the regional composition of the consumer market play in determining the optimal integration strategies. In the process, we identify three distinct "complementarities" that link firms' foreign investment decisions for different stages of production.

JEL Classification: F23, F12, L22

Keywords: direct foreign investment, multinational corporations, intra-firm trade, vertical integration.

*We acknowledge with thanks the support of the National Science Foundation (SES 9904480 and SES 0211748) and the US-Israel Binational Science Foundation (2002132).

1 Introduction

The globalization process of recent years has been expressed in the growth of many types of international transactions, but few more salient than the expansion in the activity of multinational firms. The growth rate of sales by foreign affiliates of multinational corporations outpaced the growth of exports of goods and non-factor services by almost seven percent per year from 1990 to 2001. Gross product by all foreign affiliates accounted for an estimated eleven percent of world GDP in 2001, while exports by these affiliates represented an estimated 35 percent of total world trade (UNCTAD, 2002).

Multinational firms have pursued a multitude of strategies for international expansion, as described in the *World Investment Report* (UNCTAD, 1998) and cited by Yeaple (2003). Firms have opened foreign affiliates to perform activities ranging from R&D to after-sales service, and including production of parts and components, assembly, and wholesale and retail distribution, among others. Some firms procure parts from subsidiaries in many countries and assemble them in a single location. Others concentrate production of parts in one place and assemble final products in several plants located close to their customers. Still others erect an integrated plant in a low-wage country and use it to serve consumers around the globe. The motives for foreign direct investment (FDI) are similarly diverse, but the potential for factor-cost savings, for transportation-cost and trading-cost savings, and for the realization of economies of scale seem to be among the primary inducements.

The theory of international trade and foreign direct investment traditionally has distinguished two forms of multinational activity based on alternative reasons why a firm might opt to locate production or other activities abroad (see, for example, Markusen [2002, pp.17-20]). *Vertical* multinationals are firms that geographically separate various stages of production. Such fragmentation of the production process typically is motivated by cost considerations arising from cross-country differences in factor prices. For example, Helpman (1984) and Helpman and Krugman (1985) model multinational firms that maintain their headquarters in one country but manufacture elsewhere so as to conserve on production costs. In contrast, *horizontal* multinationals are firms that replicate most or all of the production process in several locations. These multi-plant firms often are motivated by potential savings of transport and trading costs. In the models developed by Markusen (1984), Brainard (1997) and Markusen and Venables (1998, 2000), for example, firms with headquarters in a home country produce final output in plants that serve consumers in each of two national markets.

The distinction between vertical and horizontal FDI is clear enough when there are two countries and two production activities, namely headquarter operations and "manufacturing." But with more countries and more stages of production, some organizational forms do not fit neatly into either of these categories. For example, a multinational firm might manufacture goods in a foreign subsidiary and sell the output primarily in third-country markets; Ekholm et al. (2003) term such activity "export-platform FDI." Or a firm might perform intermediate

stages of production in one country to save on production costs and subsequent stages in several plants to conserve on transport costs. Yeaple (2003) follows the *World Investment Report* in referring to this as a "complex integration strategy." Feinberg and Keane (2003) report that, in their sample of U.S. multinationals with affiliates in Canada, only 12 percent of the firms have negligible intra-firm flows of intermediate goods and thus can be considered to be purely horizontal multinationals, while only 19 percent of the firms have intra-firm flows of intermediate goods in only one direction, which would make them purely vertical multinationals. The remaining 69 percent of firms are what they call "hybrids"; i.e., firms that are pursuing more complex integration strategies. Similarly, Hanson et al. (2001) describe the rich patterns of FDI they find in their data pertaining to operations by U.S. multinationals and their foreign affiliates. They document and analyze the roles played by foreign affiliates as export platforms, as producers adding value to inputs acquired from their U.S. parents, and as wholesale distributors in foreign markets. Based on their analysis of data for the 1990's, Hanson et al. conclude that "the literature's benchmark distinction between horizontal and vertical FDI does not capture the range of strategies that multinationals use."

Both Yeaple (2003) and Ekholm et al. (2003) examine theoretically the determinants of firms' choices among a limited set of integration strategies that includes an option for FDI that is neither purely horizontal nor purely vertical. Yeaple studies a model with two identical "Northern" countries and a third, "Southern" country in which firms headquartered in one of the Northern countries need two produced inputs to assemble differentiated final goods. One component can be produced more cheaply in the North, the other in the South. Shipping entails an "iceberg" transport cost that is a similar proportion of output for intermediate goods as for final goods. All consumption of the differentiated final goods takes place in the North. In this context, Yeaple compares the profitability of four integration strategies: (i) a "national firm" that produces both of the components in the same Northern country as where its headquarters are located; (ii) a "vertical multinational" that produces one component in the South and the other in the firm's home country; (ii) a "horizontal multinational" that maintains integrated production facilities (that produce both components) in both Northern countries, and (iv) a "complex multinational" that produces one component in the South and the other in both Northern countries. In Yeaple's model of symmetric producers, all firms adopt the same integration strategy in equilibrium. Yeaple shows how the viability of the four different organizational forms depends on factor-price differentials, shipping costs, and the fixed costs of establishing subsidiaries in the North and South.

Ekholm et al. (2003) also study a setting with two similar Northern countries and a single Southern country. Theirs is a duopoly model, with one firm headquartered in each country in the North. Each of these firms must produce an intermediate good in its home country but may assemble final output in one or more plants located in any or all of the countries. Thus, each firm chooses among four options: (i) a national firm that conducts all activities at

home, (ii) a purely horizontal multinational that assembles in both Northern countries; (iii) a pure export platform, with all assembly in the South; and (iv) a hybrid multinational, with assembly in both the home country and the South. Like Yeaple, Ekholm at al. examine how the organizational choices reflect transport costs, the relative cost advantage of the South, and the fixed costs associated with foreign investment.

Our concerns in this paper are somewhat similar to those of Yeaple (2003) and Ekholm et al. (2003), but we aim to shed light on the determinants of integration strategy when firms face a richer array of choices. Our goal is to provide a reasonably general analysis in which a variety of different complex integration strategies can emerge in equilibrium. In our model, as with the others, there are three countries; namely, two, symmetric Northern countries that we call "East" and "West" and a low-wage country that we call "South." The firms that produce differentiated products must perform two production activities besides their headquarter services; they first must produce intermediate goods and then must assemble these goods into a final product. Either production of intermediate goods, or assembly, or both may be separated geographically from a firm's headquarters, and a firm may perform these activities in one or several locations.

We assume that the cost of producing components and of assembly are lower in the South than in the North. A firm must bear a fixed cost for each plant it operates abroad to produce intermediate goods and a (possibly different) fixed cost for each foreign subsidiary that assembles final goods. Both intermediate goods and final goods may be costly to trade, and the cost of transporting the two types of goods (relative to the value of output) may differ. The key parameters that we use to describe an industry are the sizes of the transport costs for intermediate and final goods, the relative size of the fixed costs for different types of subsidiaries, and the share of the consumer market that resides in the South.

We also allow for heterogeneity among the firms in an industry. Following Melitz (2002) and Helpman et al. (2004), we assume that each entrant into an industry draws a productivity level from a known distribution. By the time that firms make their decisions about integration strategy, they have learned about their own potential productivity levels. In equilibrium, firms with different productivity levels may make different choices about their organizational form. Thus, our model can account for the coexistence of a variety of forms in the same industry, in keeping with the evidence reported by Hanson et al. (2001) and Feinberg and Keane (2003).

A main theme that we stress throughout the paper is that important complementarities link a firm's decisions about where to locate its various activities. Yeaple (2003) was the first to point out one such complementarity. Here, we elaborate on his observation and distinguish three different forms of complementarity. A "unit-cost complementarity" arises when a firm locates one production activity in a low-wage country and thereby achieves a lower unit cost. With a lower cost, the firm will wish to produce a greater volume of

output and so will have greater incentive to shift other production activities to the low-wage venue. A "source-of-components" complementarity operates for an intermediate range of transport costs for final goods. When the elasticity of substitution between different production activities is not too high, the proportional savings that can be generated by reducing the cost of one activity is greater when the cost of the activity is lower. Then, for an intermediate range of transport costs, it will be profitable to move assembly operations to the low-wage country only if intermediate goods also are produced at low cost. Finally, an "agglomeration complementarity" always exists when intermediate goods are costly to transport, because firms then have an incentive to locate their production of these goods near to their assembly operations.

In our model with heterogeneous firms, we are able to show how the complementarities are reflected in the response of the fraction of firms that choose a given integration strategy with foreign investment in one activity to changes in the cost of conducting the other activity abroad. Both the unit-cost complementarity and the agglomeration complementarity imply that in industries with higher fixed costs of FDI in intermediate goods, there should be a lower share of firms engaged in assembly abroad. In addition, the source-of-components complementarity implies that for an intermediate range of transport costs of final goods, higher fixed costs of FDI in components are associated with a higher fraction of firms that perform assembly in the home country, or more generally, in the North. These implications of the analysis can be subjected to empirical scrutiny.

The remainder of this paper is organized as follows. In Section 2, we develop our model of firms that must choose where to produce intermediate goods and where to assemble final products. The firms in an industry share similar fixed costs of opening foreign subsidiaries, similar costs of shipping components, and similar costs of shipping final goods. They face symmetric demands but differ in their potential productivity. In Section 3, we analyze the equilibrium integration strategies that emerge in the absence of transport costs. In this simple case we are able to develop intuition about the sorting of firms by productivity level and show how the parameters describing fixed costs and the relative size of the South affect the choices of organizational form. We are also able to isolate the unit-cost complementarity, which is present even when transport costs are nil. Section 4 introduces transportation costs for final goods and consider the full range of possible costs from low to high. Again we examine how different parameters describing industry conditions color the equilibrium choices by firms with different productivity levels and we show how a source-of-components complementarity arises for an intermediate range of shipping costs for final goods. Section 5 contains a discussion of some interesting cases that arise when intermediate goods too are costly to transport. Such costs give rise to an agglomeration complementarity, which is discussed in this secion. Section 6 concludes.

2 The Model

We develop a simple model in which firms face a choice between performing activities at home and engaging in foreign direct investment (FDI) to conserve on either production or trading costs. We distinguish between "assembly activities"—those that result in a finished product ready for sale to consumers —and "intermediate activities"—those that can be performed in any location so long as the output later is transported to the place of assembly. In our model, there are three countries and two stages of production. Following Ekholm et al. (2003) and Yeaple (2003), we assume that one of the countries ('South') has low production costs and a relatively small market for the goods produced by the integrated firms, while the other two ('East' and 'West', together comprising the 'North') have larger markets and higher wages, and are fully symmetric.

Households consume goods produced by $J+1$ industries. One industry supplies a homogeneous good under competitive conditions. The others manufacture differentiated products. Consumers share similar preferences that can be represented by the utility function

$$U = x_0 + \sum_{j=1}^{J} \frac{1}{\mu_j \alpha_j^{\alpha_j}} X_j^{\mu_j}, \qquad 0 < \mu_j < 1, \tag{1}$$

where x_0 is consumption of the homogeneous good and X_j is an index of consumption of the differentiated outputs of industry $j \in \{1, \ldots, J\}$. The consumption index for industry j is a CES aggregate of the amounts consumed of the different varieties. That is,

$$X_j = \left[\int_0^{n_j} x_j(i)^{\alpha_j} di \right]^{1/\alpha_j}, \;\; 0 < \alpha_j < 1, \tag{2}$$

where $x_j(i)$ is consumption of the i^{th} variety of industry j and n_j is the measure (number) of varieties in that industry. With this utility function, the elasticity of substitution between any pair of goods produced by industry j is $1/(1-\alpha_j)$. We assume that $\alpha_j > \mu_j$, so that the brands in a given industry substitute more closely for one another than they do for the outputs of a different industry.

We distinguish the countries in several ways. First, firms in the North are more productive than those in the South in producing the homogeneous good. This creates a gap between Northern and Southern wages. We assume that one unit of labor is needed to produce one unit of the homogenous good in East or West, but that $1/w > 1$ units of labor are needed to produce one unit of the good in South. We also assume that the homogeneous good is produced in equilibrium in all three countries and take this good to be the numeraire. Then $w^E = w^W = 1 > w^S = w$, where w^ℓ is the wage in country ℓ. Second, the sizes of the markets for differentiated products may differ; we denote by M^ℓ the number of households in country ℓ that consume differentiated products and assume that $M^E = M^W = M^N > M^S$.

Finally, we assume that firms can enter as producers of differentiated products only in the two Northern countries and that such firms must locate their headquarters in their country of origin.

Entry into industry j requires h_j units of local labor in East or West. With this fee, an entrant acquires the design for a differentiated product and learns its productivity level. Productivity levels in industry j are independent draws from a cumulative distribution function, $G_j(\theta)$. A firm in industry j with productivity θ produces final output according to the production function $\theta F_j(m, a)$, where m is the quantity of a specialized, intermediate input and a is the level of assembly activity. The intermediate goods can be produced in a different location from the assembly activity, but if so, the intermediates must be shipped to the place of assembly before a final good can be produced. The location of assembly determines the (pre-shipment) location of the final good.

We take $F_j(\cdot)$ to be an increasing and concave function with constant returns to scale and an elasticity of substitution between m and a no greater than one. Let $c_j(p_m, p_a)$ denote the unit cost function dual to $F_j(m, a)$, where p_i is the effective price of input i in the place of assembly (including delivery costs). Then $c_j(p_m, p_a)/\theta$ is the per-unit variable cost of production in this location for a firm with productivity θ.

A firm in industry j that produces its intermediate inputs in a different country from that in which its headquarters are located bears an extra (fixed) cost of g_j units of home labor for communication and governance. These costs are the same for a firm that produces the intermediates in the other Northern country as for one that produces them in the South. Similarly, a firm that engages in FDI in assembly incurs extra fixed costs of f_j units of home labor. Iceberg transportation costs may apply to both intermediate inputs and final goods. Specifically, a firm in industry j must ship $\tau_j \geq 1$ units of the intermediate good to deliver one unit of the good to a distant place of assembly and $t_j \geq 1$ units of the final good to deliver one unit of the good to a distant place of consumption.

We assume that the manufacture of one unit of an intermediate good requires one unit of local labor in the place of production and that one unit of assembly activity requires one unit of local labor in the place of assembly. With these assumptions, the South enjoys a comparative advantage both in assembly and in production of intermediate goods, relative to production of the homogeneous good x_0.[1]

It is now straightforward to calculate the variable cost to a firm in industry j of delivering one unit of the final good to a given market by means of alternative integration strategies. Consider for example a firm in East with productivity θ that wishes to deliver final goods to consumers in West. Such a firm would pay $t_j c_j(1, 1)/\theta$ per unit to produce and assemble the

[1] We have also examined situations with different production structures that admit a comparative advantage for the South in one of the activities undertaken by the integrated firms. For small comparative advantage in one of these activities, our results are unaffected. Larger degrees of comparative advantage modify our result in fairly intuitive ways.

Table 1: Fixed and Per-Unit Variable Costs

production m	assembly a	fixed cost	per-unit variable cost
in H	in H	0	$c(1,1)/\theta$
in H	in S	f	$c(1,w)/\theta$
in S	in H	g	$c(w,1)/\theta$
in S	in S	$f+g$	$c(w,w)/\theta$

good at home (including the cost of shipping to West), whereas it would pay $t_j c_j(w,w)/\theta$ per unit to conduct all production and assembly activity in South. Still another possibility would be to produce intermediates in South and perform assembly in West, thereby avoiding the transport cost for final goods. The variable cost associated with this strategy would be $c_j(\tau_j w, 1)/\theta$ per unit, considering the cost of shipping the intermediates from South to West.

3 Zero Transport Costs

We begin our analysis with the case of costless international transport. It is useful to examine this simple case, because it highlights the trade-off between the fixed costs of FDI and the variable-cost savings that can be achieved by performing certain activities in the low-wage South (as in Helpman et al. [2004]) and the complementarities that exist between FDI decisions for different stages of development (as in Yeaple [2003]).

In what follows, we consider firms in a particular industry j and omit the subscript j from the variables and parameters of interest. We focus on the variation in productivity levels across firms in the industry, as indexed by θ. A firm may have its headquarters in East or West. Since these two countries are fully symmetric, it is more convenient to refer to H, the home country of the firm in question, and R, the "other" Northern country in which the firm will sell its output. This means, of course, that if $H = E$, $R = W$; and if $H = W$, $R = E$.

With costless shipping, an integrated firm with headquarters in H never opts to perform any activity in country R, because the variable costs are the same there as at home and FDI would impose extra fixed costs. Moreover, a firm has no reason to undertake a given activity in multiple locations, because this would impose additional governance costs without conserving on any transport costs. Thus, only four integration strategies remain for consideration with costless trade: production of intermediates might take place either in H or S and assembly might occur either in H or S. Table 1 shows the fixed and per-unit variable costs associated with each of these strategies. The fixed costs indicated are those *extra* costs that result from operating one or more foreign subsidiaries.

The first row depicts a strategy of home production. With this strategy, the firm serves the foreign markets in R and S with exports from its home assembly plant. As is clear, this

strategy minimizes the fixed costs of governance, but provides a relatively high per-unit cost, because factor prices are higher in E or W than in S. The following two rows depict strategies of "partial globalization"; either intermediates are produced at home and assembled in South (second row), or vice versa (third row). These strategies yield intermediate levels of fixed and variable costs; they cannot be ranked vis-à-vis one another without further information about the cost function $c(\cdot)$ and the sizes of the fixed costs for the two types of foreign subsidiaries. With assembly in S, the firm exports intermediates from its home plant, and then exports finished goods from S to consumers in H and R. This means that the strategy combines elements of "vertical FDI" and what Ekholm et al. (2003) have termed "export-platform FDI." With intermediates produced in S, there again is intra-firm trade, as well as exports of final goods from H to markets in R and S. The bottom row depicts a strategy of complete globalization, whereby all production activities are performed in the low-wage South. Here, fixed costs are highest, variable costs are lowest, and the markets in H and R are served by exports from South. With this strategy, there is no trade in intermediate goods.

We can readily compare the operating profits that a firm with productivity θ can achieve under the alternative strategies. Considering the form of consumer preferences in (1) and (2), every firm in the industry faces a demand function in market ℓ given by

$$x^\ell = \alpha^{-\alpha/(1-\alpha)} M^\ell \left(X^\ell\right)^{(\mu-\alpha)/(1-\alpha)} \left(p^\ell\right)^{-1/(1-\alpha)} , \tag{3}$$

where X^ℓ is the aggregate consumption index for varieties in the industry in country ℓ and p^ℓ is the price it charges there. Each producer treats the aggregate consumption indexes as given. Therefore, it maximizes profits by charging a price in each market that is a multiple $1/\alpha$ of its per-unit variable cost of serving that market. Since the per-unit cost of serving every market is the same when transport costs are zero, so too are the optimal prices associated with a given strategy. It follows from the demand function in (3) that, for any strategy with an extra fixed cost of k and a per-unit variable cost of c/θ, the maximum attainable operating profits are

$$\pi = (1-\alpha)\bar{Y}\Theta c^{-\alpha/(1-\alpha)} - k,$$

where $\Theta \equiv \theta^{\alpha/(1-\alpha)}$ is a transformed measure of the firm's productivity and $\bar{Y} \equiv \sum M^\ell \left(X^\ell\right)^{(\mu-\alpha)/(1-\alpha)}$ is a measure of world demand.

In Figure 1, we depict the operating profits attainable from home production (the top row in Table 1) and complete globalization in South (the bottom row in Table 1), for different levels of productivity Θ. These profits, which we denote by $\pi_{H,H}$ and $\pi_{S,S}$, are given by

$$\pi_{H,H} = \frac{(1-\alpha)\bar{Y}\Theta}{C(1,1)} \tag{4}$$

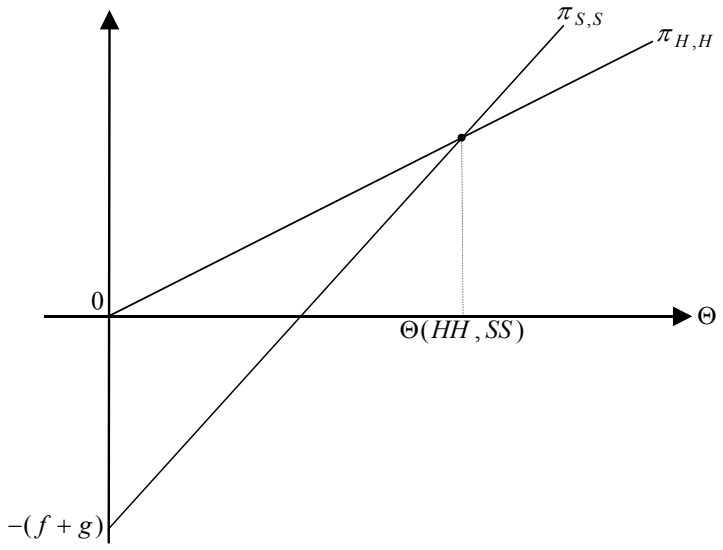

Figure 1: Profitability of home production and complete globalization

and

$$\pi_{S,S} = \frac{(1-\alpha)\bar{Y}\Theta}{C(w,w)} - (f+g) \tag{5}$$

respectively, where $C(p_m, p_a) \equiv [c(p_m, p_a)]^{\alpha/(1-\alpha)}$ is a transformed measure of unit cost. The figure shows that firms with low productivity prefer home production whereas firms with high productivity prefer FDI, in keeping with the findings of Helpman et al. (2004). The reason, of course, is that FDI offers the prospect of lower per-unit costs and the potential variable cost savings are most valuable to productive firms that anticipate producing high volumes of output.

Next consider the firm's option to locate only its assembly operations in South. The potential operating profits from this integration strategy for a firm with productivity Θ are

$$\pi_{H,S} = \frac{(1-\alpha)\bar{Y}\Theta}{C(1,w)} - f \ . \tag{6}$$

If we were to add $\pi_{H,S}$ to Figure 1, it would have an intercept between those of $\pi_{H,H}$ and $\pi_{S,S}$ and a slope steeper than $\pi_{H,H}$ but less steep than $\pi_{S,S}$. Thus, if locating only assembly in South is to be viable at any level of productivity, this strategy must be at least as profitable as concentrating both activities in either location at the productivity level labelled $\Theta(HH, SS)$ in the figure. But this requires[2]

$$\frac{g}{f} \geq \gamma_H \equiv \frac{C(1,1)}{C(w,w)} \left[\frac{C(1,w) - C(w,w)}{C(1,1) - C(1,w)} \right] \ . \tag{7}$$

[2] To derive this condition, we calculate $\Theta(HH, SS)$ as the value of Θ that equates $\pi_{H,H}$ and $\pi_{S,S}$, and then compare $\pi_{H,S}$ and $\pi_{H,H}$ at $\Theta = \Theta(HH, SS)$.

Leaving this strategy aside for the moment, the firm also has the option to produce intermediate goods in South and assemble final goods at home. This strategy offers a firm with productivity Θ operating profits of

$$\pi_{S,H} = \frac{(1-\alpha)\bar{Y}\Theta}{C(w,1)} - g \ . \tag{8}$$

Again, the intercept and slope are intermediate between those for the two lines shown in Figure 1, and the viability of the strategy at any Θ requires that it be at least as profitable as the other two at $\Theta = \Theta(HH, SS)$. This in turn requires

$$\frac{g}{f} \leq \gamma_L \equiv \frac{C(w,w)}{C(1,1)} \left[\frac{C(1,1) - C(w,1)}{C(w,1) - C(w,w)} \right] \ . \tag{9}$$

From (7) and (9) we conclude that if

$$\gamma_L < \frac{g}{f} < \gamma_H \ ,$$

all firms will concentrate their production activities in either H or S. Our assumption that the elasticity of substitution between intermediates and assembly in the production of final goods is no greater than one ensures that the upper limit in this string of inequalities exceeds the lower limit.[3] It follows that there always exists a range of values of g/f for which partial globalization is not optimal for any firm.

Suppose now that the fixed costs of operating a foreign assembly plant are small relative to the fixed costs of operating a foreign plant to manufacture intermediate goods; i.e., $g/f > \gamma_H$. Then a firm with productivity level at or near $\Theta(HH, SS)$ prefers to locate its assembly in South and manufacture intermediates at home to any other integration strategy. Figure 2 shows the operating profits $\pi_{H,S}$ (as well as $\pi_{H,H}$ and $\pi_{S,S}$) for this case. Clearly, firms with productivity below $\Theta(HH, HS)$ conduct all operations at home, firms with intermediate productivity level between $\Theta(HH, HS)$ and $\Theta(HS, SS)$ conduct their intermediate production at home and their assembly operations in South, and firms with productivity above $\Theta(HS, SS)$ perform all of their production activities in South.

[3] It can be shown that $\gamma_H > \gamma_L$ if and only if

$$\frac{1}{C(w,w)} + \frac{1}{C(1,1)} > \frac{1}{C(w,1)} + \frac{1}{C(1,w)};$$

i.e., if and only if the function $1/C(\cdot)$ is supermodular. But $1/C(p_m, p_a) \equiv [c(p_m, p_a)]^{\alpha/(1-\alpha)}$ is supermodular if it is twice differentiable and

$$\frac{c(p_m,p_a)\left[\partial^2 c(p_m,p_a)/\partial p_m \partial p_a\right]}{[\partial c(p_m,p_a)/\partial p_m][\partial c(p_m,p_a)/\partial p_a]} < \frac{1}{1-\alpha}.$$

The left-hand side of this inequality is the elasticity of substitution between m and a in the production of final goods, which is no greater than one by assumption. Therefore, the inequality holds for all positive values of α.

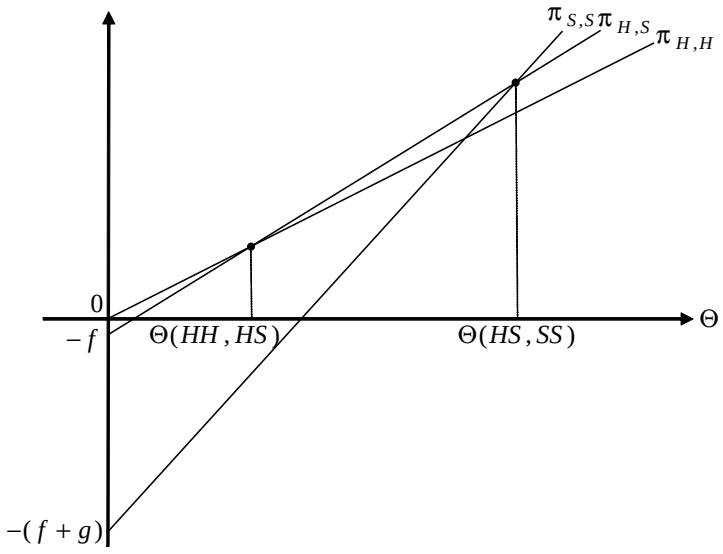

Figure 2: Partial globalization optimal for intermediate productivity levels

The case in which the fixed costs of FDI in assembly are large relative to the fixed costs of FDI in intermediates is qualitatively similar. With g/f small enough so that $g/f < \gamma_L$, the line representing $\pi_{S,H}$ will cut $\pi_{H,H}$ at some relatively low productivity level $\Theta(HH, SH)$ that is to the left of $\Theta(HH, SS)$ in Figure 1, and will cut $\pi_{S,S}$ at some relatively high productivity level $\Theta(SH, SS)$ to the right of $\Theta(HH, SS)$ in the figure. Then firms with productivity between $\Theta(HH, SH)$ and $\Theta(SH, SS)$ will choose to produce their intermediates in the low-wage South while conducting assembly at home.

Figure 3 shows combinations of productivity Θ and fixed costs of FDI in intermediate goods g that generate different integration strategies. The heavy lines (both solid and broken) represent boundaries between regions with different optimal strategies. In the region $\{H, H\}$ all production activity takes place in the home country; in $\{S, H\}$ intermediates are produced in South while assembly is performed at home; and so on. The figure applies for a particular value of the fixed costs of FDI in assembly f. When f changes, the boundaries between the regions shift. The appendix provides details on the construction of these boundaries. Here we illustrate the construction of two such boundaries: the broken vertical line between $\{H, H\}$ and $\{H, S\}$ and the solid, upward-sloping line between $\{H, S\}$ and $\{S, S\}$; others are constructed similarly.

The boundary between $\{H, H\}$ and $\{H, S\}$ is defined by $\pi_{H,H} = \pi_{H,S}$; these are points at which the operating profits from concentrating production in the home country are just equal to the operating profits from producing intermediates in the home country and assembling

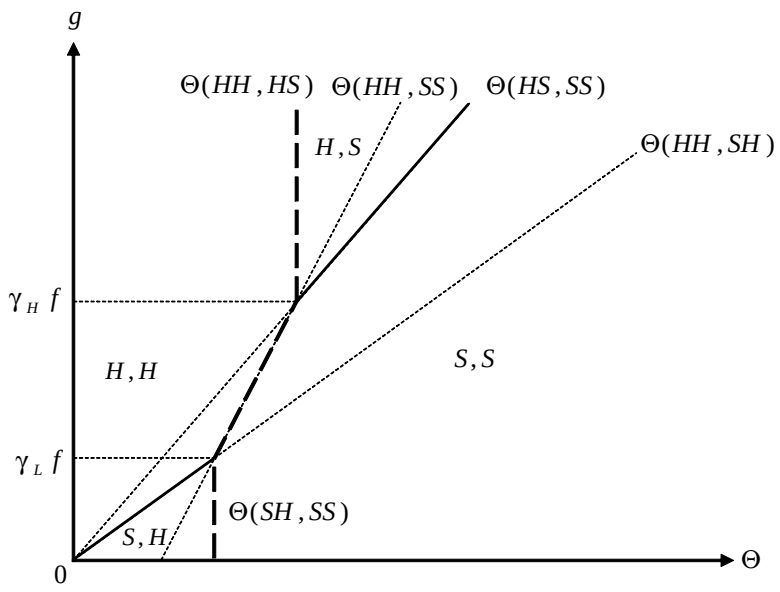

Figure 3: Integration strategies for different productivities and fixed costs of FDI in components

final goods in South. Equations (4) and (6) imply that

$$\Theta\left(HH,HS\right)=\frac{f}{\left(1-\alpha\right)\bar{Y}\left[\frac{1}{C(1,w)}-\frac{1}{C(1,1)}\right]}.$$

Clearly, the productivity level at which $\{H,H\}$ and $\{H,S\}$ yield similar operating profits does not depend on the fixed costs g of FDI in intermediate production, since neither of these strategies entails any such foreign production of components. Therefore, the boundary $\{H,H\}$ and $\{H,S\}$ is vertical line as shown in the figure. From Figure 2 we know that when $g>\gamma_H f$, $\{H,S\}$ is the optimal strategy for firms with an intermediate range of productivity levels. But at a productivity level $\Theta\left(HS,SS\right)$ defined by $\pi_{H,S}=\pi_{S,S}$, a firm will be indifferent between investing in foreign production of intermediate goods and producing its components at home. The solid boundary line in the figure is given by

$$\Theta\left(HS,SS\right)=\frac{g}{\left(1-\alpha\right)\bar{Y}\left[\frac{1}{C(w,w)}-\frac{1}{C(1,w)}\right]},$$

which is represented by a ray through the origin. Here, the higher are the fixed costs g of FDI in intermediate production, the higher must be a firm's productivity level before it would choose to invest in component production in South.

Figure 3 shows that, for all strictly positive values of g, firms with low productivity perform all production activities in their home country and export their final product to R

and S. These firms intend to produce relatively little output, so the savings in variable cost offered by FDI does not justify the higher fixed costs of FDI. Firms with intermediate levels of productivity may separate their production of intermediates from their assembly operations, depending on the size of g. If so, such firms will engage in intra-firm trade in addition to exporting final output either from their home assembly plant or from an export platform in South. Finally, high-productivity firms perform all operations in the low-wage South so as to take greatest advantage of the low per-unit costs there.[4]

Our analysis can be used to highlight an important complementarity that generally exists between the decisions to invest abroad at different stages of production. Note that FDI in assembly takes place to the right of the heavy broken lines in Figure 3. Firms with productivity less than $\Theta(SH, SS)$ do not engage in FDI in assembly no matter what is the size of g, while firms with productivity greater than $\Theta(HH, HS)$ do engage in FDI for all values of g. But for firms with intermediate productivity levels such that $\Theta(SH, SS) < \Theta < \Theta(HH, HS)$, FDI in assembly will be profitable only if the fixed costs of FDI in component production is low. In other words, for these firms it is profitable either to shift all production activities to South, or to shift none.[5] We shall refer to this complementarity as a "unit-cost complementarity"; it arises from the fact that when a firm invests in performing any activity in the low-cost region, such FDI reduces its unit cost, which raises desired output, and thus increases the return to performing other production activities in the low-cost region.

We can readily compute the fraction of firms that choose each of the alternative integration strategies. It follows immediately from our discussion that, when the unit-cost complementarity operates (as it does when g lies between $\gamma_L f$ and $\gamma_H f$), the fraction of firms that engage in FDI in assembly rises as the fixed cost of investment in intermediate production falls. Similarly, the fraction of firms that invest in foreign production of intermediate goods rises as the fixed costs of FDI in assembly fall. In this sense, decisions about the location of one stage of production are linked to those about the location of the other.

[4] The model can be closed to construct an industry equilibrium, which determines the aggregate consumption index X. Define the envelope of the profit functions as

$$\pi(\Theta) = \max_{z_1 \in \{H,S\}, z_2 \in \{H,S\}} \pi_{z_1,z_2}(\Theta) \ ,$$

where $\pi(\Theta)$ is the operating profit earned by a firm with productivity Θ when it pursues its optimal integration strategy. Given the distribution of productivity levels $G(\theta)$, the free-entry condition can be written as

$$\int_0^\infty \pi \left[\theta^{\alpha/(1-\alpha)}\right] dG(\theta) = h \ .$$

Since the profit function is increasing in the measure of world demand $\bar{Y}$, which in turn is increasing in the aggregate consumption index X, the free-entry condition uniquely determines the industry value for X. All other industry variables, including the number of varieties and the cut-off points for each integration strategy can now be computed using this value of X.

[5] Yeaple (2003) makes a similar point about cost complementarity in the decisions of a single firm.

4 Transport Costs for Final Goods

In this section, we allow for costly transport of final goods while maintaining the assumption that intermediates can be shipped costlessly. For example, the intermediates may represent services that can be performed remotely and then moved electronically.

We shall find that the optimal integration strategies vary with the size of the transport costs. We begin with a case in which transport costs for final goods are reasonably small; in particular, we suppose that

$$1 < t < \frac{c(1,1)}{c(1,w)} \ . \tag{10}$$

When inequality (10) is satisfied, the variable cost of serving any market is minimized by assembly in South, no matter where the intermediate goods are produced. To see this, observe first that if the intermediates are produced in H or R, the cost of serving any market from an assembly plant in the North is at least $c(1,1)$. But this exceeds the cost of serving the same market from South, which is at most $tc(1,w)$. Next observe that if intermediates are produced in South, the per-unit variable cost of serving any market from an assembly plant in the North is at least $c(w,1)$, while the per-unit cost of serving the same market from a plant in South is at most $tc(w,w)$. However, $c(w,1)/c(w,w) > c(1,1)/c(1,w)$[6], so inequality (10) ensures that $c(w,1) > tc(w,w)$ as well.

Under these circumstances, a firm with headquarters in H will not conduct any activity in R. Intermediate goods are no less costly to produce in R than in H and can be shipped costlessly from one to the other. By producing these goods in R, the firm would needlessly incur the extra fixed costs of FDI. And if assembly is to be conducted outside of H, the delivered cost of serving any market from S are lower than the cost of serving the market from R, while the fixed costs of an assembly plant are the same in the two locations.

We can also rule out any integration strategy in which a given activity is performed in more than one location. If it is worthwhile for the firm to bear the fixed costs of opening a facility to manufacture intermediate goods in South, the firm produces all of its intermediates there to take full advantage of the low production costs. The same is true for assembly, considering the reasonably low cost of shipping goods. It follows that each firm chooses one of four integration strategies; these are the same set of strategies that we considered in Section 3.

A firm's decision calculus is similar to that described in Section 3, except that now it must take into account the relative size of the market in South when deciding whether to open facilities there. We define $Y^\ell \equiv M^\ell (X^\ell)^{(\mu-\alpha)/(1-\alpha)}$ as a measure of market size in country ℓ and $\sigma \equiv Y^S/\bar{Y}$ as the share of the South in world demand for industry output.

It is now straightforward to show that the four regions of the optimal integration strategies

[6]Note that $c(1,1)/c(1,w) < c(1,w)/c(w,w)$ if and only if $\log c(1,1) + \log c(w,w) < \log c(1,w) + \log c(w,1)$; i.e., if and only if $\log c(p_m, p_a)$ is submodular. But $\log c(p_m, p_a)$ indeed is submodular when the elasticity of substitution between m and a is less than one, because $\partial^2 \log c(p_m, p_a)/\partial p_m \partial p_a < 1$.

are as depicted in Figure 3, except that now the parameters γ_L and γ_H and the boundaries between regions depend on σ, the relative size of South. This means that, as in Section 3, there is a unit-cost complementarity between the two forms of FDI. In particular, the higher is the fixed costs of FDI in components the smaller is the fraction of firms that invest in assembly in the South. And similarly, the higher are the fixed costs of FDI in assembly, the smaller is the fraction of firms that invest in components in the South. Now, however, the fraction of firms that invest in assembly in South also depends on the relative size of South. As one would expect, for given fixed costs of FDI in components and assembly, the larger is the relative size of the South, the larger is the fraction of firms that invests in assembly there.[7]

Next we consider an industry with moderate transport costs such that

$$\frac{c(1,1)}{c(1,w)} < t < \frac{c(w,1)}{c(w,w)} \ . \tag{11}$$

When transport costs fall in this intermediate range, a market in the North is served at lower per-unit cost by exports from the South than by local assembly if and only if the intermediate goods are also produced in the South. The fact that $c(w,1)/c(w,w) > c(1,1)/c(1,w)$ introduces a second source of complementarity between the two forms of FDI, distinct from the unit-cost complementarity that we identified before. The inequality implies that the potential cost savings from conducting assembly in a low-cost region is relatively greater when components are also produced there. We refer to this as a "source-of-components complementarity".

Again, it is never optimal for a firm with its headquarters in H to produce intermediate goods in R. Such a firm could instead produce the intermediate goods in S and achieve lower variable costs while incurring the same fixed costs. Also, a firm has no reason to produce intermediate goods in two locations, because these goods are costless to transport. Thus, all of the integration strategies that might be viable in this case involve production of intermediates either in H or in S (but not both).

A firm that chooses to produce its intermediate goods in H will serve its home market with final goods that have been assembled there as well, in view of the left-most inequality in (11). Also, a firm that chooses to produce its intermediate goods in S will either perform all of its assembly there or else assemble all final goods at home. With intermediate goods from the South, assembly in South offers the lowest variable cost of serving any market in view of the right-most inequality in (11). Thus, a firm that elects to bear the fixed costs of FDI in assembly will serve all markets from there. But a firm may choose to avoid the fixed costs of FDI in assembly by performing its assembly at home. We are left with six integration strategies to consider when transport costs are moderate: Southern production

[7] See the appendix for details.

of intermediate goods with assembly either in H or in S; or home production of intermediate goods with assembly in H, in H and S, in H and R, or in H, S and R.

Let us begin once again, by considering the operating profits that a firm with productivity Θ can achieve by concentrating all production activities either in H or in S. By performing all activities at home, the firm avoids all fixed costs of FDI but bears a very high per-unit cost of $tc(1,1)$ of serving the markets in R and S, and a reasonably high per-unit cost of $c(1,1)$ of serving the home market. Nonetheless, this strategy will be attractive to firms with very low productivity, because these firms intend to produce low volumes of output. The associated operating profits are given by

$$\pi_{H,H} = (1-\alpha)\bar{Y}\Theta\frac{[(\frac{1-\sigma}{2})(1+T)+\sigma]}{TC(1,1)} \ ,$$

where $T = t^{\alpha/(1-\alpha)}$ is another measure of transport costs. At the other extreme, by performing all activities in South, a firm pays a high total fixed cost of $f+g$, but it attains the lowest possible per-unit cost of serving each of the markets. Operating profits then are given by

$$\pi_{S,S} = (1-\alpha)\bar{Y}\Theta\frac{[(1-\sigma)+\sigma T]}{TC(w,w)} - (f+g) \ . \tag{12}$$

Such a strategy will appeal to firms with high productivity that intend to produce great volumes of output. It follows, as before, that the lowest productivity firms in an industry concentrate their activities in the home country and the highest productivity firms perform all production activities in the low-wage South.

Next consider a strategy that involves production of intermediate goods in the home country and assembly in H and in at least one other country. If assembly takes place only in H and R, the firm is engaged in horizontal FDI to conserve on shipping costs to the other Northern market. The resulting profits are[8]

$$\pi_{H,HR}(\Theta) = (1-\alpha)\bar{Y}\Theta\frac{[(1-\sigma)T+\sigma]}{TC(1,1)} - f \ . \tag{13}$$

If assembly takes place only in H and S, the firm uses its plant in S both to serve the Southern market and as an export platform for sales to R. Then operating profits are given by

$$\pi_{H,HS}(\Theta) = (1-\alpha)\bar{Y}\Theta\left[\frac{\frac{1-\sigma}{2}}{C(1,1)} + \frac{\frac{1-\sigma}{2}+\sigma T}{TC(1,w)}\right] - f \ . \tag{14}$$

Finally, if assembly takes place in all three countries, each market is served by products

[8] In this notation, the subscript on π gives the index of the country (or countries) in which the firm produces its intermediates followed by a comma and then a list of the countries in which assembly takes place.

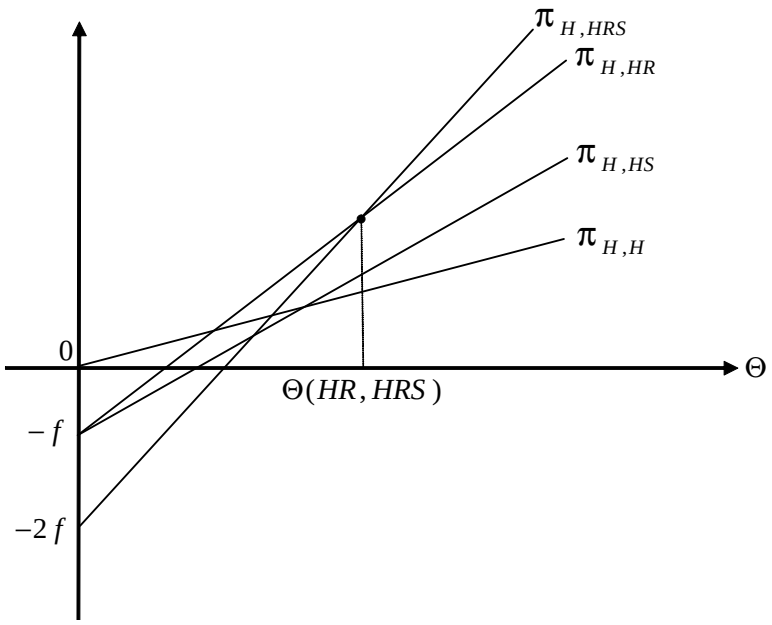

Figure 4: Assembly in multiple plants with moderate transport costs

assembled locally, and operating profits are given by

$$\pi_{H,HRS}(\Theta) = (1-\alpha)\bar{Y}\Theta\left[\frac{1-\sigma}{C(1,1)} + \frac{\sigma}{C(1,w)}\right] - 2f \ . \tag{15}$$

Figure 4 depicts the operating profits for the integration strategies that involve assembly in more than one location. Of the three, the strategy in which the firm operates assembly plants in all three countries has the highest total fixed costs and the lowest per-unit variable cost. The variable costs are low with this strategy, because the firm avoids all shipping costs. The strategy is preferred to the other two by firms with relatively high productivity. The remaining two strategies with assembly in H and one other location entail similar fixed costs of FDI. The figure shows a case in which a strategy of assembling in S for sales in S and R generates higher variable costs and therefore lower operating profits than a strategy of assembling in R for these markets.[9] This case applies whenever the market share of the South is smaller than $\hat{\sigma}_H$, where

$$\hat{\sigma}_H = \frac{TC(1,w) - C(1,1)}{(2T-1)C(1,1) + (T-2)C(1,w)} \tag{16}$$

is the critical value of σ at which it is equally profitable to assemble in H and R as it is to assemble in H and S, when intermediate goods are produced in H. If $\sigma > \hat{\sigma}_H$, then

[9] Equivalently, the firm might assemble in R for sales in R and serve the market in S with exports from H. Once the fixed cost of an assembly plant in R has been borne, the cost of exporting to S from R or H are the same.

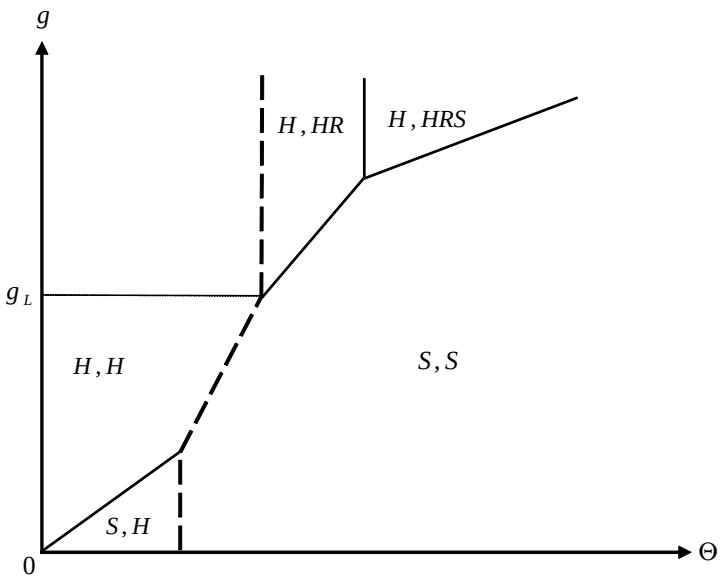

Figure 5: Integration strategies for moderate transport costs and $\sigma < \hat{\sigma}_H$

$\pi_{H,HR} < \pi_{H,HS}$ for all $\Theta > 0$.[10]

Figure 4 also shows the operating profits that a firm would earn by concentrating all activity at home. From the figure, it is apparent how firms would locate their assembly operations (as a function of their productivity level), conditional on their having decided to produce intermediate goods at home. Those with low productivity prefer a single assembly plant at home to any other assembly pattern, while those with high productivity prefer to have assembly plants in all three countries. The firms with intermediate levels of productivity prefer to have an assembly operation at home and in one other country; in the South if σ is relatively large, and in R otherwise.

Finally, we must consider each firm's option to produce its intermediate goods in South and then assemble final goods in either H or S. If intermediate goods are produced in South and assembly takes place at home, operating profits are

$$\pi_{S,H} = (1-\alpha)\bar{Y}\Theta\left[\frac{(1-\sigma)(1+T)+2\sigma}{2TC(w,1)}\right] - g,$$

whereas if all production activities take place in South the profits are given by the expression in (12). Among these two strategies, firms with low productivity prefer the former and firms with high productivity prefer the latter.

Figure 5 depicts the optimal integration strategies as functions of the fixed costs of FDI in intermediate goods g and the firm-level productivity parameter Θ, for given fixed costs of

[10] Our restrictions on transport costs imply that $\hat{\sigma}_H < 1/3$. That is, this critical value of the relative size of South requires the South to be smaller than a typical Northern country.

FDI in assembly, moderate transport costs, and a relatively small South (i.e., $\sigma < \hat{\sigma}_H$).[11] FDI in assembly takes place in regions $\{S, S\}$, $\{H, HR\}$ and $\{H, HRS\}$, i.e., to the right of the heavy broken lines. However, the form and function of the foreign investment varies across these different regions. In $\{S, S\}$, final goods are assembled only in South, which serves as an export platform to the two Northern countries. In $\{H, HR\}$ assembly takes place in the two Northern countries and FDI in R is used to serve the market in R alone. Finally, in $\{H, HRS\}$ assembly takes place in all three countries. In this case, FDI in assembly eliminates all trade in final goods.

It is clear from this figure that the fraction of firms that engage in FDI in assembly, undistinguished by form and purpose, rises as the fixed costs of FDI in components falls; that is, the unit-cost complementarity that we identified for low transport costs continues to operate. The interesting new feature is that FDI in assembly now may take place in different countries and the source-of-components complementarity affects the attractiveness of the alternative locations differently. Whereas the fraction of firms that conducts assembly in South rises (or does not change) as the cost of FDI in components falls, the fraction that invests in assembly in the other Northern country actually falls (or does not change) when the fixed costs of FDI in components fall. When g is large, the fraction of firms that conducts some assembly in South is invariant to the size of fixed costs for FDI in components. But the composition of firms with assembly operations in South does change with g, as a reduction in g expands the fraction that invests only in South and reduces the fraction that invests in both S and R.

The shift in the composition of FDI in assembly that takes place when g is above g_L in Figure 5 reflects the aforementioned source-of-components complementarity. Recall that when transportation costs are moderate, a market in the North can be served at lower per-unit cost by exports from the South than by local assembly if and only if the intermediate goods are also produced in the South. Small fixed costs of FDI in components encourage production of components in the South. As a result, some of the lower productivity firms that otherwise would prefer to produce their components in the home country will opt to produce them in the South as g falls. For these firms, it also becomes more profitable to assemble final goods in South, rather than in R. Thus, as g falls in the range where $g > g_L$, the fraction of firms that produces components and assemble final goods in South rises while the fraction that produces components at home and assemble final goods in East and West falls.

[11] The construction of Figure 5 is explained in the appendix. In our working paper, Grossman, Helpman and Szeidl (2003), we also derive the optimal integration strategies for cases in which $\sigma > \hat{\sigma}_H$. When the South is relatively large, the region with assembly in H and R does not exist; instead, there is one with assembly in H and S.

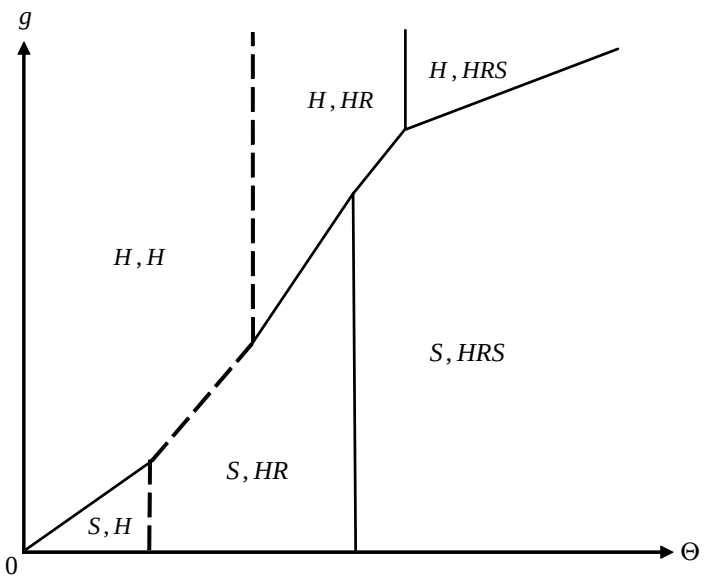

Figure 6: Integration strategies for high transport costs and $\sigma < \hat{\sigma}_Z$

Finally, we consider an industry in which shipping final goods is quite costly, so that

$$t > \frac{c(w,1)}{c(w,w)} \ . \tag{17}$$

In such circumstances, the lowest variable cost of serving *any* market is achieved by local assembly.[12] Figure 6 depicts the optimal integration strategies for industries with such high transport costs. In drawing the figure, the fixed costs of FDI in assembly f and aggregate income $\bar{Y}$ are held constant; we also draw a case in which the relative size of the South is small.[13]

The difference between the optimal integration strategies with high and moderate transport costs can be seen by comparing Figures 6 and 5. The main difference is that high transport costs encourage firms to conduct assembly in R. In particular, whereas when t is moderate firms that produce components abroad engage in foreign assembly, if at all, only in South, when t is large such firms may choose to conduct foreign assembly also in R, or perhaps only in R. We also find that the regions with foreign assembly of intermediate goods

[12] Recall that an elasticity of substitution between intermediate goods and assembly smaller than one ensures that $c(w,1)/c(w,w) > c(1,1)/c(1,w)$. Therefore, when (17) is satisfied, $tc(1,w) > c(1,1)$.

[13] We show in the appendix that a configuration of regions similar to that in Figure 6 applies whenever $\sigma < \hat{\sigma}_H$; the only possible variations are that the boundary between $\{S, HR\}$ and $\{S, HRS\}$ may be located to the left of the boundary between $\{H, H\}$ and $\{H, HR\}$ for some parameter values; and the region in which $\{S, HR\}$ is the optimal strategy may not exist at all. We show in our working paper that when $\sigma > \hat{\sigma}_H$, the region in which the optimal strategy is $\{H, HR\}$ is replaced by one in which the optimal strategy is $\{H, HS\}$; and for even larger values of σ, the region in which the optimal strategy is $\{S, HR\}$ is replaced by one in which the optimal strategy is $\{S, HS\}$.

manufactured at home expand in size.

As should be familiar by now, the fraction of firms that invest in assembly in foreign countries rises when the fixed costs of FDI in components decline. FDI in assembly takes place in regions $\{S, HR\}$, $\{S, HRS\}$, $\{H, HR\}$ and $\{H, HRS\}$, to the right of the broken heavy line. The composition of firms that invest in foreign assembly also changes as g falls. The fraction of firms that produce components in South and assemble them in the two Northern countries (only) rises gradually from zero (once g is low enough) and then becomes constant. Note also that for all g such that some firms produce components in South and assemble them in East and West only, there are also higher productivity firms that produce components in South and assemble them in all three countries. The fraction of the latter type of firms rises as g falls and then becomes constant. For very high g, the total fraction of firms that assemble final goods in some foreign country is invariant to g, but the composition of this fraction changes with g. In particular, the fraction of firms that produce components in the South and assemble final goods in all three countries rises as g falls, whereas the fraction of firms that produce components in the home country and conduct assembly in all three countries declines as g falls. At such high levels of g the fraction of firms that produce components in H and assemble final goods in East and West is constant.

To summarize, our model predicts an increasing share of firms that engage in FDI in assembly as the fixed costs of FDI in components fall. This qualitative prediction does not depend on the size of shipping costs for final goods. And our model predicts a relationship between the size of the fixed costs of FDI in components and the composition of FDI in assembly that depends on the size of this transport cost. In particular, for moderate transport costs of final goods, we have identified a source-of-components complementarity that induces a negative correlation between the fixed costs of FDI in components and the fraction of firms that invest in assembly in foreign countries.

5 Transport Costs for Intermediate Goods

Up until now, we have assumed that intermediate goods can be moved costlessly to any place of assembly. This simplifying assumption allowed us to examine how variations in the cost of transporting final goods, in relative market size, and in the relative fixed costs of FDI in different activities affect firms' decisions about global integration.

In this section, we introduce a cost of trading intermediate inputs (i.e., $\tau > 1$). To avoid a detailed taxonomy, however, we explore only cases in which the cost of transporting intermediate goods is high and South is negligible in size ($\sigma = 0$). Under these conditions, firms have no incentive to locate their assembly operations in S as a means to serve the Southern market. Rather, if a firm opens an assembly plant in South, it is because it wishes to use its plant there as an export platform. We focus attention on cases when τ is sufficiently

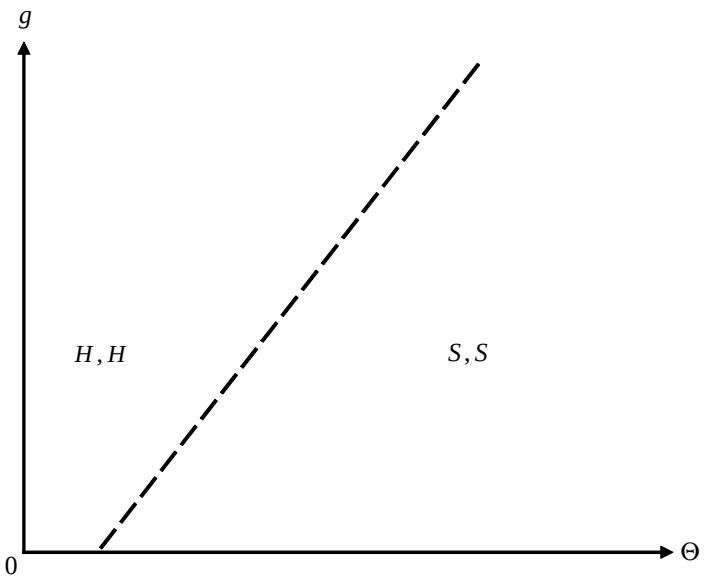

Figure 7: Integration strategies for high transport costs of intermediate goods and no transport costs of final goods

large to satisfy $c(\tau, 1)/c(1, 1) > 1/w$; i.e., the cost premium from producing intermediates in R and shipping them for assembly in H relative to the cost of concentrating all production in H exceeds the cost premium from producing intermediates in a Northern country relative to producing them in South.[14] By examining this case, we are able to identify clearly yet another complementarity between the two forms of FDI.

We first consider the case in which there are no transport costs for final goods. When $t = 1$, there can be no source-of-components complementarity. In this cae, only two integration strategies may be viable: a firm either concentrates the production of intermediate goods and the assembly of final goods in its home country or else it concentrates these activities in the South. To see why this is so, note that in the absence of transport costs of final goods FDI in assembly in the other Northern country is never optimal, because it is cheaper to assemble final goods in South and ship them to the North than it is to assemble them in R. And FDI in components can be profitable only if a firm also invests in foreign assembly.[15] So either a firm conducts all production activities in South or else it keeps all activities at home.

The case in hand points to another complementarity between FDI in production of components and FDI in assembly, namely an "agglomeration complementarity." It arises because

[14] In our working paper, Grossman, Helpman and Szeidl (2003) , we discuss the optimal integration strategies for other possible sizes of transport costs for intermediate goods. See also the appendix, which provides the details of the following analysis.

[15] The assumption that $c(\tau, 1)/c(1, 1) > 1/w$ implies that $wc(\tau, 1) > c(1, 1)$, which in turn implies that $c(\tau w, 1) > c(1, 1)$.

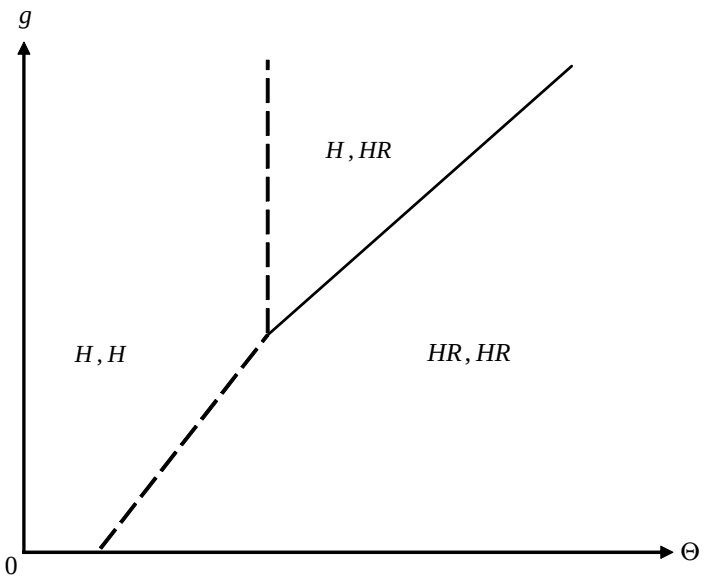

Figure 8: Integration strategies for high transport costs of final and intermediate goods

when intermediate goods are costly to ship firms have an incentive to assemble the final products close to their facility for producing intermediate goods. The point is seen most clearly when, as here, the final goods are costless to ship, so that there is no offsetting incentive based on shipping costs to locate assembly near to consumers.

The optimal integration strategies for the case of high shipping costs for intermediate goods and zero shipping costs for final goods are shown in Figure 7. As before, it is the high-productivity firms (for any given f and g) that will find it worthwhile to incur the fixed costs of foreign investment. The upward sloping boundary between the two regions implies, once again, that the fraction of firms that engage in FDI in assembly rises when the fixed costs of FDI in production of intermediate goods declines. Here the complementarity between the two forms of FDI is present at every level of fixed costs g. This reflects the fact that the agglomeration complementarity is present for all g, when τ is sufficiently high and $t = 1$.

One additional case worth mentioning arises when final goods also are costly to transport and in fact sufficiently so that $t > c(\tau, 1)/c(1, 1) > 1/w$. Under such conditions, it never pays to assemble in the South. But since the agglomeration complementarity is still present, if assembly never occurs in the South, neither does production of intermediate goods take place there. We show in our working paper, Grossman, Helpman and Szeidl (2003), that three integration strategies are viable: production of intermediates and assembly of final goods may be concentrated at home; production of intermediate goods may take place at home with assembly in each Northern market; or intermediate goods may be produced in each Northern market for assembly in a nearby location and sale to local consumers. Figure 8 shows the values of Θ and g for which each strategy is optimal, given world income $\bar{Y}$ and

the size of fixed costs for FDI in assembly f. [16]

We see that the fraction of firms that invests in assembly is invariant to the size of fixed costs for FDI in components when g is sufficiently high, but it varies inversely with the size of these fixed costs when g is relatively small. Again, the agglomeration complementarity is reflected in a co-movement in the two forms of FDI.

6 Conclusions

In this paper, we have examined the joint determination of international trade and foreign direct investment in a setting in which firms may choose among a rich array of integration strategies. In our analysis, firms that are headquartered in a Northern country supply differentiated final goods to two national markets in the North and one in the South. Each such firm must produce an intermediate input and conduct assembly activities in order to generate a final product. The firms may produce intermediate goods in their home country, in the other Northern country, or in the South. Similarly, assembly may take place in any of the three locations. And firms may choose to maintain plants for either or both stages of production in multiple locations. Accordingly, there are many possible organizational forms available to firms. Each firm's choice has implications for the pattern of trade in intermediate and final goods.

We characterized industries by the sizes of the fixed costs of maintaining a foreign subsidiary for production of intermediate goods and for assembly, the costs of transporting intermediate and final goods internationally, and the fraction of the consumer demand that resides in the low-wage South. For each industry, we derived the equilibrium organizational forms for the heterogeneous firms in the industry that differ in their productivity levels.

In an industry in which transportation of intermediate and final goods is costless, the relative size of the fixed costs for foreign investment in intermediate goods and assembly determines the set of organizational forms that are observed in equilibrium. Here, the relative sizes of the markets have no bearing on the equilibrium choices and there is no intra-industry FDI. Firms with low productivity choose an integration strategy that minimize the fixed costs of operation, whereas firms with high productivity seek to minimize the variable costs of serving the various markets. A unit-cost complementarity links a firm's decisions about foreign investment; if circumstances lead a firm to conduct one production activity in the low-wage South, the firm will have lower variable costs (compared to when it conducts this activity at home), thus a higher optimal volume of output and a greater incentive to shift the other activity to the low-wage country as well.

When final goods are costly to transport, the set of integration strategies that are used in an industry depends on the size of these shipping costs. For small transport costs, no

[16] When t is very close to $c(\tau,1)/c(1,1)$, region H, HR disappears.

single production activity takes place in multiple locations and each activity is performed either in a firm's home country or in the South. For higher transport costs, some firms in an industry that produce their intermediate goods at home will choose to assemble them both in the home country and the other Northern country, while others may conduct assembly in all three locations. Finally, when the costs of transporting final goods are very high, there will be some firms that produce intermediate goods in the South that choose to assemble these goods near to their Northern markets. A source-of-components complementarity exists for an intermediate range of transport costs. For shipping costs in this range, the unit-cost savings from conducting assembly in the low-wage South can justify the extra cost of shipping from there only if the intermediate goods also are produced at low unit cost. The presence of this complementarity implies a response of the composition of FDI in assembly to changes in the cost of FDI in components; as the fixed costs of FDI in components fall, the fraction of firms that produce their intermediate goods in the South increases, of course, but then the fraction of firms that performs assembly in the South rises at the expense of the fraction that assembles in multiple Northern locations.

Finally, costly transport of intermediate goods can make it attractive for a firm to produce intermediate goods in multiple locations. An agglomeration complementarity exists, because a firm that locates an assembly operation abroad will have an incentive to produce components nearby in order to avoid the cost of moving the intermediate goods. When the cost of shipping intermediate goods (as a fraction of value) is high but that of shipping final goods is less so, a fall in the fixed costs of either form of FDI leads to an increase in the fraction of firms that operate integrated production facilities in the South. When the costs of shipping both intermediate and final goods are large, a fall in the fixed costs of either form of FDI is associated instead with an increase in the fraction of firms that operate integrated production facilities in both Northern countries.

One limitation of our analysis in this paper is that we take the boundaries of the firm as given. That is, we have simply assumed that firms must produce their own intermediate goods and perform assembly in-house. In other recent work (Grossman and Helpman, 2003, 2004a, 2004b) two of us have studied how contracting problems interact with factor-price differentials and transport costs to determine which activities are outsourced and which performed within a firms' corporate boundaries. In those papers, the range of strategies open to the multinational firm was substantially narrower than here. Ultimately, we would like a theory that simultaneously explains the make-or-buy decision and the organization of the multinational firm. Such a theory could help explain the broad range of corporate strategies that are found in the firm-level data.

References

[1] Axtell, Robert L. (2001), "Zipf Distribution of U.S. Firm Sizes," *Science,* 293, 1818-1820.

[2] Brainard, Lael S. (1997) "An Empirical Assessment of the Proximity-Concentration Trade-off between Multinational Sales and Trade," *American Economic Review*, 87, 520-544.

[3] Ekholm, Karolina, Forslid, Rikard and Markusen, James (2003) "Export Platform Foreign Direct Investment" NBER Working Paper No. 9517.

[4] Feinberg, Susan E. and Keane, Michael P. (2003) "Accounting for the Growth of MNC-Based Trade Using a Structural Model of U.S. MNCs" University of Maryland, manuscript.

[5] Grossman, Gene M, and Helpman, Elhanan (2003), "Outsourcing versus FDI in Industry Equilibrium," *Journal of the European Economic Association*, 1, 317-327.

[6] Grossman Gene M. and Helpman, Elhanan (2004a), "Outsourcing in a Global Economy," forthcoming in the *Review of Economic Studies.*

[7] Grossman Gene M. and Helpman, Elhanan (2004b), "Managerial Incentives and the International Organization of Production," forthcoming in the *Journal of International Economics.*

[8] Grossman, Gene M., Helpman, Elhanan, and Szeidl, Adam (2003), "Optimal Integration Strategies for the Multinational Firm," NBER Working Paper No. 10189.

[9] Hanson, Gordon H., Mataloni, Raymond J. and Slaughter, Matthew J. (2001), "Expansion Strategies of U.S. Multinational Corporations," *Brookings Trade Forum 2001*, 245-294.

[10] Helpman, Elhanan (1984) "A Simple Theory of International Trade with Multinational Corporations," *Journal of Political Economy*, 92, 451-471.

[11] Helpman, Elhanan and Krugman, Paul (1985) *Market Structure and Foreign Trade,* Cambridge, MA, The MIT Press.

[12] Helpman, Elhanan, Melitz, Marc J. and Yeaple, Stephen R. (2004). "Export versus FDI with Hetrogeneous Firms," *American Economic Review,*.94, 300-316.

[13] Markusen, James R. (1984) "Multinationals, Multi-Plant Economies, and the Gains from Trade," *Journal of International Economics,* 16, 205-226.

[14] Markusen, James R. (2002) *Multinational Firms and the Theory of International Trade*, Cambridge, MA and London, UK: The MIT Press.

[15] Markusen, James R. and Venables, Anthony J. (1998) "Multinational Firms and the New Trade Theory," *Journal of International Economics*, 46, 183-203.

[16] Markusen, James R. and Venables, Anthony J. (2000) "The Theory of Endowment, Intra-Industry, and Multinational Trade," *Journal of International Economics*, 52, 209–234.

[17] Melitz, Marc J. (2002) "The Impact of Trade on Intra-Industry Reallocations on Aggregate Industry Productivity," NBER Working Paper No. 8881, forthcoming in *Econometrica*.

[18] UNCTAD (1998) *World Investment Report: Trends and Determinants*, New York and Geneva: United Nations Conference on Trade and Development.

[19] UNCTAD (2002) *World Investment Report: Transnational Corporations and Export Competitiveness*, New York and Geneva: United Nations Conference on Trade and Development.

[20] Yeaple, Stephen R. (2003), "The Complex Integration Strategies of Multinationals and Cross Country Dependencies in the Structure of Foreign Direct Investment," *Journal of International Economics*, 60, 293-314.

Appendix

DERIVATIONS FOR SECTION 3

It follows from equations (4) and (5) that the profit lines $\pi_{H,H}$ and $\pi_{S,S}$ intersect at the productivity level

$$\Theta(HH,SS) = \frac{f+g}{(1-\alpha)\bar{Y}} \cdot \frac{C(w,w)C(1,1)}{C(1,1)-C(w,w)}.$$

Then

$$\pi_{H,H}[\Theta(HH,SS)] = (f+g) \cdot \frac{C(w,w)}{C(1,1)-C(w,w)}$$

and

$$\pi_{H,S}[\Theta(HH,SS)] = (f+g) \cdot \frac{C(1,1)}{C(1,w)} \frac{C(w,w)}{C(1,1)-C(w,w)} - f \ ,$$

as well as

$$\pi_{S,H}[\Theta(HH,SS)] = (f+g) \cdot \frac{C(1,1)}{C(w,1)} \frac{C(w,w)}{C(1,1)-C(w,w)} - g$$

from (6) and (8). It follows that $\pi_{H,S}[\Theta(HH,SS)] > \pi_{H,H}[\Theta(HH,SS)]$ that is, locating only assembly in the South is viable at some productivity levels, if and only if (7) holds. Likewise, $\pi_{S,H}[\Theta(HH,SS)] > \pi_{H,H}[\Theta(HH,SS)]$, that is, locating only intermediate goods production in the South is viable at some productivity levels, if and only if (9) holds.

From $\pi_{H,H} = \pi_{H,S}$ and equations (4) and (6) we have

$$\Theta(HH,HS) = \frac{f}{(1-\alpha)\bar{Y}} \cdot \frac{C(1,w)C(1,1)}{C(1,1)-C(1,w)} \ .$$

We can derive in similar fashion

$$\begin{aligned}
\Theta(HH,SH) &= \frac{g}{(1-\alpha)\bar{Y}} \cdot \frac{C(w,1)C(1,1)}{C(1,1)-C(w,1)} \ , \\
\Theta(SH,SS) &= \frac{f}{(1-\alpha)\bar{Y}} \cdot \frac{C(w,1)C(w,w)}{C(w,1)-C(w,w)} \ , \\
\Theta(HS,SS) &= \frac{g}{(1-\alpha)\bar{Y}} \cdot \frac{C(1,w)C(w,w)}{C(1,w)-C(w,w)} \ .
\end{aligned}$$

To understand the construction of Figure 3, note that for $g < \gamma_L f$ the viable strategies are $\{H,H\}$, $\{S,H\}$ and $\{S,S\}$, with $\{H,H\}$ being optimal for low, $\{S,H\}$ for intermediate, and $\{S,S\}$ for high productivity levels. Therefore, the boundaries between these three regions are given by $\Theta(HH,SH)$ and by $\Theta(SH,SS)$. As shown above, $\Theta(SH,SS)$ does not depend on g, thus the corresponding boundary is vertical in Figure 3. On the other hand, $\Theta(HH,SH)$ is proportional to g, which explains why the corresponding boundary lies on a ray from the origin. By definition, $\Theta(HH,SH) = \Theta(SH,SS)$ when $\pi_{H,H} = \pi_{S,H} = \pi_{S,S}$, that is, when $g = \gamma_L f$.

For $\gamma_L f < g < \gamma_H f$, there are only two viable strategies, $\{H, H\}$ and $\{S, S\}$. From the above expression for $\Theta(HH, SS)$, the corresponding boundary is given by the equation

$$g = (1-\alpha)\bar{Y}\frac{C(1,1) - C(w,w)}{C(w,w)C(1,1)} \cdot \Theta - f$$

which is an upward sloping line, as depicted in Figure 3. Clearly, this line also passes through the point where $\pi_{H,H} = \pi_{S,H} = \pi_{S,S}$. Finally, the range where $\gamma_H f < g$ is explained in the main text.

DERIVATIONS FOR SECTION 4

Low transport costs

The profit functions $\pi_{H,H}$ and $\pi_{S,S}$ now become

$$\pi_{H,H} = \frac{(1-\alpha)\Theta}{TC(1,1)}\left[Y^N(1+T) + Y^S\right] = (1-\alpha)\bar{Y}\Theta\frac{\left[\frac{1-\sigma}{2}(1+T)+\sigma\right]}{TC(1,1)}$$

and

$$\pi_{S,S} = \frac{(1-\alpha)\Theta}{TC(w,w)}\left[2Y^N + TY^S\right] - (f+g) = (1-\alpha)\bar{Y}\Theta\frac{[1-\sigma+T\sigma]}{TC(w,w)} - (f+g).$$

Equating them yields

$$\Theta(HH, SS) = \frac{f+g}{(1-\alpha)\bar{Y}} \cdot \left[\frac{1-\sigma+T\sigma}{TC(w,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(1,1)}\right]^{-1}.$$

Now

$$\pi_{H,H}[\Theta(HH, SS)] = \frac{(f+g)}{TC(1,1)} \cdot \left[\frac{1-\sigma+T\sigma}{TC(w,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(1,1)}\right]^{-1} \cdot \left[\frac{1-\sigma}{2}(1+T)+\sigma\right]$$

and

$$\pi_{H,S}[\Theta(HH, SS)] = \frac{f+g}{TC(1,w)} \cdot \left[\frac{1-\sigma+T\sigma}{TC(w,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(1,1)}\right]^{-1}[1-\sigma+\sigma T] - f.$$

Therefore $\pi_{H,S}[\Theta(HH, SS)] > \pi_{H,H}[\Theta(HH, SS)]$ if and only if

$$\frac{g}{f} > \gamma_H = \frac{C(1,1)}{C(w,w)}\left[\frac{C(1,w) - C(w,w)}{C(1,1) - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{1-\sigma+\sigma T}C(1,w)}\right].$$

Likewise,

$$\pi_{S,H}[\Theta(HH,SS)] = \frac{f+g}{TC(w,1)} \cdot \left[\frac{1-\sigma+T\sigma}{TC(w,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(1,1)}\right]^{-1} \left[\frac{1-\sigma}{2}(1+T)+\sigma\right] - g$$

and $\pi_{S,H}[\Theta(HH,SS)] > \pi_{H,H}[\Theta(HH,SS)]$ if and only if

$$\frac{g}{f} < \gamma_L = \frac{C(w,w)}{C(1,1)} \left[\frac{C(1,1)-C(w,1)}{C(w,1)\frac{1-\sigma+T\sigma}{\frac{1-\sigma}{2}(1+T)+\sigma} - C(w,w)}\right].$$

We can derive $\Theta(HH,HS)$ from $\pi_{H,H} = \pi_{H,S}$, which yields

$$\Theta(HH,HS) = \frac{f}{(1-\alpha)\bar{Y}} \cdot \left[\frac{1-\sigma+\sigma T}{TC(1,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(1,1)}\right]^{-1}.$$

Similarly, we find that

$$\begin{aligned}
\Theta(HS,SS) &= \frac{g}{(1-\alpha)\bar{Y}} \cdot \frac{1}{1-\sigma+T\sigma} \cdot \left[\frac{1}{TC(w,w)} - \frac{1}{TC(1,w)}\right]^{-1}, \\
\Theta(HH,SH) &= \frac{g}{(1-\alpha)\bar{Y}} \cdot \frac{1}{\frac{1-\sigma}{2}(1+T)+\sigma} \cdot \left[\frac{1}{TC(w,1)} - \frac{1}{TC(1,1)}\right]^{-1}, \\
\Theta(SH,SS) &= \frac{f}{(1-\alpha)\bar{Y}} \cdot \left[\frac{1-\sigma+\sigma T}{TC(w,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(w,1)}\right]^{-1}.
\end{aligned}$$

As in the case with $t = 1$, we have that $\Theta(HH,HS)$ and $\Theta(SH,SS)$ do not depend on g, thus the corresponding lines are vertical. The rest of the construction of the figure is exactly the same as before. It is not difficult to show that $\Theta(SH,SS)$, $\Theta(HH,SS)$ and $\Theta(HH,HS)$ are all decreasing in σ, because $T > 1$. This implies that as the relative size of the South increases, a larger fraction of firms invest in assembly in the South.

Moderate transport costs

Equating $\pi_{H,HR}$ and $\pi_{H,HS}$ we obtain

$$(1-\alpha)\bar{Y}\Theta\frac{[(1-\sigma)T+\sigma]}{TC(1,1)} - f = (1-\alpha)\bar{Y}\Theta\left[\frac{\frac{1-\sigma}{2}}{C(1,1)} + \frac{[\frac{1-\sigma}{2}+T\sigma}{TC(1,w)}\right] - f,$$

or equivalently,

$$\hat{\sigma}_H = \frac{TC(1,w)-C(1,1)}{(2T-1)C(1,1)+(T-2)C(1,w)}.$$

We now turn to explain Figure 5. First of all, note that the location of the profit functions $\pi_{H,HR}$ and $\pi_{H,HRS}$ in Figure 4 does not vary with g, because the slopes are independent of

fixed costs, and the intercepts depend only on f. The same is true for the profit function $\pi_{H,H}$, which is not shown in that Figure. This implies that the cutoff productivity values $\Theta(H, HR)$ and $\Theta(HR, HRS)$ do not depend on g, and the corresponding boundaries in Figure 5 are vertical lines. One can easily show that

$$\Theta(H, HR) = \frac{f}{(1-\alpha)\bar{Y}} \cdot \frac{2TC(1,1)}{(1-\sigma)(T-1)}$$

and

$$\Theta(HR, HRS) = \frac{f}{(1-\alpha)\bar{Y}} \cdot \frac{1}{\sigma} \cdot \frac{TC(1,w)C(1,1)}{TC(1,1) - C(1,w)}.$$

Therefore, $\Theta(HR, HRS) > \Theta(H, HR)$ if and only if

$$(1-\sigma)(T-1) > 2\sigma TC(1,1)\left[\frac{1}{C(1,w)} - \frac{1}{TC(1,1)}\right].$$

Since both sides are linear in σ, this inequality will hold for the set of parameter values we are interested in if and only if it is true for $\sigma = 0$ and $\sigma = \hat{\sigma}_H$. The inequality is clearly satisfied for $\sigma = 0$. For $\sigma = \hat{\sigma}_H$, substituting in for $\hat{\sigma}_H$ yields

$$T - 1 > \frac{TC(1,w) - C(1,1)}{TC(1,1) - C(1,w)} TC(1,1)\left[\frac{1}{C(1,w)} - \frac{1}{TC(1,1)}\right] = T - \frac{C(1,1)}{C(1,w)}$$

which holds. It follows that $\Theta(HR, HRS) > \Theta(H, HR)$.

Next consider the strategies $\{S, H\}$ and $\{S, S\}$. First, note that as g increases, the profit functions $\pi_{S,H}$ and $\pi_{S,S}$ shift down in parallel fashion in Figure 4, because their intercepts contain a term g. This implies that the cutoff productivity value $\Theta(SH, SS)$ does not depend on g. This value was explicitly calculated above to be

$$\Theta(SH, SS) = \frac{f}{(1-\alpha)\bar{Y}} \cdot \left[\frac{1-\sigma+\sigma T}{TC(w,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(w,1)}\right]^{-1}.$$

It is easy to show that $\Theta(SH, SS) < \Theta(H, HR)$ if and only if

$$\frac{1-\sigma+\sigma T}{TC(w,w)} - \frac{\frac{1-\sigma}{2}(1+T)+\sigma}{TC(w,1)} > \frac{(1-\sigma)(T-1)}{2TC(1,1)}.$$

Since both sides are linear in σ, this inequality will always hold if it is true for $\sigma = 0$ and $\sigma = 1$. The inequality is obvious for $\sigma = 1$. For $\sigma = 0$, it is equivalent to

$$\frac{1}{TC(w,w)} - \frac{1+T}{2TC(w,1)} > \frac{T-1}{2TC(1,1)}.$$

Because of the bound $TC(w,w) < C(w,1)$, this inequality will be true if

$$\frac{1}{C(w,1)} - \frac{1+T}{2TC(w,1)} > \frac{T-1}{2TC(1,1)} \ ,$$

that is, when

$$C(1,1) > C(w,1)$$

which is satisfied.

It is now easy to show that Figure 5 describes the optimal integration strategies. For g very low, that is, as long as the profit level $\pi_{S,H}[\Theta(SH,SS)]$ is above $\pi_{H,H}$, the upper envelope of $\pi_{S,H}$ and $\pi_{S,S}$ will be above $\pi_{H,HR}$ and $\pi_{H,HRS}$. This is because at $\Theta(SH,SS)$ the strategy $\{S,S\}$ dominates $\{H,H\}$, but at this point $\{H,H\}$ still dominates $\{H,HR\}$ and $\{H,HRS\}$ (since $\Theta(SH,SS) < \Theta(H,HR) < \Theta(H,HRS)$)– and clearly, once $\{S,S\}$ dominates for some productivity level, it will dominate for every higher productivity level too. As g increases, the profit level $\pi_{S,H}(\Theta(SH,SS)]$ falls below $\pi_{H,H}$. This means that $\{S,H\}$ becomes dominated by $\{H,H\}$ and $\{S,S\}$. At this level of g, there are no other viable strategies yet, again because $\Theta(SH,SS) < \Theta(H,HR)$. As g increases further, first $\{H,HR\}$, and then $\{H,HRS\}$ also become viable. This explains the regions plotted in Figure 5. To see why are all the boundaries straight lines, note that every formula we have is linear in g, thus so are the all the cutoff values.

High transport costs

For the sake of completeness, we discuss all possible scenarios for the range $\sigma < \hat{\sigma}_H$. Equating $\pi_{S,HR}$ and $\pi_{S,HS}$ yields

$$(1-\alpha)\bar{Y}\Theta\frac{[(1-\sigma)T+\sigma]}{TC(w,1)} - (f+g) = (1-\alpha)\bar{Y}\Theta\left[\frac{\frac{1-\sigma}{2}}{C(w,1)} + \frac{[\frac{1-\sigma}{2}+T\sigma}{TC(w,w)}\right] - (f+g) \ ,$$

which implies that (S,HS) will never be used as long as

$$\sigma < \hat{\sigma}_S = \frac{C(w,1) - TC(w,w)}{(2-T)C(w,w) + (1-2T)C(w,1)}.$$

Using the formulas for $\hat{\sigma}_H$ and $\hat{\sigma}_S$ it is easy to show that

$$\hat{\sigma}_H < \frac{1}{3} < \hat{\sigma}_S$$

which implies in particular that $\min(\hat{\sigma}_H, \hat{\sigma}_S) = \hat{\sigma}_H$.

We now turn to explain how optimal integration strategies look like. We show in our working paper, Grossman, Helpman and Szeidl (2003), that the only viable integration strategies when $\sigma < \hat{\sigma}_H$ are $\{H,H\}$, $\{H,HR\}$, $\{H,HRS\}$ and $\{S,H\}$, $\{S,HR\}$ and $\{S,HRS\}$.

Note that, like in the case with moderate transport costs, the profit lines corresponding to the first three of these strategies do not change as we vary g. Thus the cutoff productivity levels $\Theta(H, HR)$ and $\Theta(HR, HRS)$ do not depend on g, and the argument showing $\Theta(HR, HRS) > \Theta(H, HR)$ given in the moderate transport costs case continues to be valid, because we did not make use of the bounds for T.

Let us turn to strategies $\{S, H\}$, $\{S, HR\}$ and $\{S, HRS\}$. Because each of these involves setting up an intermediate production facility in the South, their profit lines all shift in parallel when we vary g. This implies that the corresponding cutoff productivity levels $\Theta[(S, H), (S, HR)]$ and $\Theta[(S, HR), (S, HRS)]$ do not depend on g. One can calculate

$$\Theta[(S, H), (S, HR)] = \frac{f}{(1-\alpha)\bar{Y}} \cdot \frac{2TC(w, 1)}{(1-\sigma)(T-1)}$$

and

$$\Theta[(S, HR), (S, HRS)] = \frac{f}{(1-\alpha)\bar{Y}} \cdot \frac{1}{\sigma} \left[\frac{1}{C(w, w)} - \frac{1}{TC(w, 1)} \right]^{-1}.$$

We now turn to pin down the order of the cutoff productivity levels. First, $\Theta[(S, H), (S, HR)] < \Theta[(H, H), (H, HR)]$ is easy to check. Next note that $\Theta[(S, HR), (S, HRS)] < \Theta[(H, HR), (H, HRS)]$ is equivalent to

$$\frac{1}{C(w, w)} - \frac{1}{C(1, w)} > \frac{1}{T} \left[\frac{1}{C(w, 1)} - \frac{1}{C(1, 1)} \right]$$

which holds for any $T > 1$ because the function $1/C(\cdot)$ is supermodular. Third, $\Theta[(S, H), (S, HR)] < \Theta[(S, HR), (S, HRS)]$ is equivalent to

$$\frac{\sigma}{C(w, w)} - \frac{\sigma}{TC(w, 1)} < \frac{(1-\sigma)(T-1)}{2TC(w, 1)}$$

or

$$\sigma < \hat{\sigma}_Z = \frac{(T-1)C(w, w)}{2TC(w, 1) - (3-T)C(w, w)}.$$

In general it may be possible that $\hat{\sigma}_Z < \hat{\sigma}_H$, or that the inequality goes the other way around (though for T high enough, $\hat{\sigma}_Z$ will be smaller). Assume first that $\sigma < \min(\hat{\sigma}_Z, \hat{\sigma}_H)$ holds. Under these circumstances, the order of the cutoff productivity levels we are interested in is either

$$\Theta[(S, H), (S, HR)] < \Theta[(H, H), (H, HR)] < \Theta[(S, HR), (S, HRS)] < \Theta[(H, HR), (H, HRS)]$$

or

$$\Theta[(S, H), (S, HR)] < \Theta[(S, HR), (S, HRS)] < \Theta[(H, H), (H, HR)] < \Theta[(H, HR), (H, HRS)].$$

In words, the order of the the middle two cutoff productivity levels depends on the particulars

of the cost function and other parameters. The exact form of Figure 6 depends slightly on the order of these productivity levels, although the intuition does not. The Figure is drawn assuming that the first chain of inequalities holds, which will be true if σ is small enough, so let us focus on that case first. Consider the upper envelope of the profit functions corresponding to $\{S, H\}$, $\{S, HR\}$ and $\{S, HRS\}$. First of all, note that at each productivity level Θ, this envelope is steeper than the upper envelope corresponding to $\{H, H\}$, $\{H, HR\}$ and $\{H, HRS\}$. Thus these two upper envelopes have a single crossing property: once a strategy that involves producing intermediates in the South is optimal for some productivity level, it will be optimal for every higher productivity level too. Let us now trace how the envelope corresponding to producing intermediates in the South moves when we vary g. When g is very low, this envelope intersects $\pi_{H,H}$ to the left of $\Theta(H, HR)$. Moreover, for small g, at the intersection point the upper envelope still coincides with $\pi_{S,H}$. This explains Figure 6 for low g.

As g increases, the upper envelope just discussed is shifted downwards. Once the intersection of $\pi_{S,H}$ and $\pi_{S,HR}$ shifts below $\pi_{H,H}$ (that is, when $\pi_{S,H}\{\Theta[(S, H), (S, HR)]\}$ falls below $\pi_{H,H}$) the strategy $\{S, H\}$ is no longer viable. As g further increases and the upper envelope shifts down to the level that it passes through the point where $\pi_{H,H}$ and $\pi_{H,HR}$ intersect, strategy $\{H, HR\}$ starts to become optimal for an intermediate range of productivity levels. This is illustrated in Figure 6 for intermediate levels of g.

For g even higher, the part of the upper envelope corresponding to $\pi_{S,HR}$ shifts entirely below $\pi_{H,H}$ and $\pi_{H,HR}$, and $\{S, HR\}$ is no longer optimal for any productivity level. The rest of the argument similar to the one given in the moderate transport costs case.

For the case where the second chain of inequalities holds, the only qualitative change in the Figure is that the boundary segment corresponding to $\Theta[(S, HR), (S, HRS)]$ appears to the left of the segment corresponding to $\Theta[(H, H), (H, HR)]$. This implies that, like in Figure 5, for an intermediate range of g values there will be only two integration strategies chosen in optimum, which in this case are $\{H, H\}$ and $\{S, HRS\}$.

Let us consider now the case where $\hat{\sigma}_Z < \sigma < \hat{\sigma}_H$. Note that this range may be empty, if $\min(\hat{\sigma}_Z, \hat{\sigma}_H) = \hat{\sigma}_H$. If it is not empty, then by $\hat{\sigma}_Z < \sigma$ we have that $\Theta[(S, H), (S, HR)] > \Theta[(S, HR), (S, HRS)]$ which implies that $\pi_{S,HR}$ is dominated by $\pi_{S,H}$ and $\pi_{S,HRS}$ for all productivity levels, thus there will not be a region corresponding to $\{S, HR\}$. Moreover, one can show that $\Theta[(S, H), (S, HR)] < \Theta[(H, H), (H, HR)]$ is equivalent to

$$\frac{1-\sigma}{2}\frac{T-1}{TC(w,1)} + \sigma\left[\frac{1}{C(w,w)} - \frac{1}{TC(w,1)}\right] > \frac{(1-\sigma)(T-1)}{TC(1,1)}$$

which is easily seen to hold for $\hat{\sigma}_Z < \sigma$. It follows that for $\hat{\sigma}_Z < \sigma < \hat{\sigma}_H$ the order of the

cutoff productivity levels is

$$\Theta[(S,H),(S,HRS)] < \Theta[(H,H),(H,HR)] < \Theta[(H,HR),(H,HRS)].$$

It follows that, the figure looks just like Figure 5, except that the $\{S, HR\}$ region is replaced by $\{S, HRS\}$, and accordingly there is no longer a boundary between $\{S, HR\}$ and $\{S, HRS\}$.

DERIVATIONS FOR SECTION 5

The profit functions for strategies $\{H,H\}$, $\{H,HR\}$, $\{H,HR\}$, $\{HR,HR\}$ and $\{S,S\}$ when $\sigma = 0$ are

$$\begin{aligned}
\pi_{H,H}(\Theta) &= (1-\alpha)\bar{Y}\Theta\frac{T+1}{2TC(1,1)} \\
\pi_{H,HR}(\Theta) &= (1-\alpha)\bar{Y}\Theta\left[\frac{1}{2C(1,1)}+\frac{1}{2C(\tau,1)}\right]-f \\
\pi_{HR,HR}(\Theta) &= (1-\alpha)\bar{Y}\Theta\frac{1}{C(1,1)}-(f+g) \\
\pi_{S,S}(\Theta) &= (1-\alpha)\bar{Y}\Theta\frac{1}{TC(w,w)}-(f+g).
\end{aligned}$$

It follows that when $t = 1$, we have

$$\Theta(HH,SS) = \frac{f+g}{(1-\alpha)\bar{Y}}\cdot\left[\frac{1}{C(w,w)}-\frac{1}{C(1,1)}\right]^{-1}.$$

Thus, in Figure 7 the only boundary is an upward sloping line, which assumes a positive value when $g = 0$.

To derive the boundaries in Figure 8, note that

$$\begin{aligned}
\Theta[HH,(HR,HR)] &= \frac{f+g}{(1-\alpha)\bar{Y}}\cdot\frac{2TC(1,1)}{T-1} \\
\Theta[HH,(H,HR)] &= \frac{f}{(1-\alpha)\bar{Y}}\cdot\left[\frac{1}{2C(\tau,1)}-\frac{1}{2TC(1,1)}\right]^{-1} \\
\Theta[(H,HR),(HR,HR)] &= \frac{g}{(1-\alpha)\bar{Y}}\cdot\left[\frac{1}{2C(1,1)}-\frac{1}{2C(\tau,1)}\right]^{-1}.
\end{aligned}$$

Thus the cutoff productivity value between strategies $\{H,H\}$ and $\{H,HR\}$ is independent of g. As g varies, the profit lines corresponding to these strategies are unchanged. For g very small, $\{H,H\}$ and $\{HR,HR\}$ will jointly dominate $\{H,HR\}$. But as g rises, the profit line corresponding to $\{HR,HR\}$ is shifted downwards, and eventually, $\{H,HR\}$ becomes viable for intermediate productivity levels. As usual, all boundaries are straight lines; moreover, the boundary between $\{H,HR\}$ and $\{HR,HR\}$ lies on a ray from the origin because $\Theta[(H,HR),(HR,HR)]$ is proportional to g.