This paper provides an integrative analysis of the drivers of vertical scope, using analytical and computational methods. I propose a model with two vertical segments (upstream and downstream), with firm populations that have heterogeneous capabilities, and an intermediate market subject to transaction costs, where firms can choose whether to be integrated or vertically specialized. By varying the level of transaction costs and changing the structure of the correlation between upstream-downstream capabilities in the industry, as well as economies of scale; learning curves; and the way in which profitability leads to capability improvement in the upstream and downstream segments, I generate numerical results to explain how vertical integration evolves over time. The results suggest that (a) without capability differences, even if transaction costs are nil, firms remain integrated; (b) differences in economies of scale in the two segments may create or dampen specialization, depending on the underlying capability heterogeneity structure; (c) transaction costs catalyze the underlying capability differences to drive scope; (d) dynamic factors, such as learning curves; returns to investment in capabilities; or limits to expansion exacerbate small, random capability differences and as such promote specialization; and finally, these dynamic factors can just by themselves lead to substantial specialization when they differ between the upstream and downstream segments. The model also provides a new rationale for “mixed governance” (concurrent use of both the market and integration), as well as for the initial period of vertical integration, followed by specialization, both regularities that are often observed in practice.

Keywords: Vertical Scope; Productive capabilities; Transaction Costs; Scale; Learning Curves
Over the last thirty years, significant progress has been made in understanding what determines vertical scope (Williamson, 1985, 1999; Shelanski & Klein, 1995). Most of this work has focused on a particular, narrow line of inquiry, which takes the transaction as the unit of analysis and examines the particular transaction costs (TC) which, on the margin, push a firm towards vertical integration as opposed to procurement through the market. However, several authors have argued that, capabilities as well as TC must be considered in explaining vertical scope (Argyres, 1996; Poppo & Zenger, 1998; Schilling & Steensma, 2002; Madhok, 2002; Leiblein & Miller, 2003; Hoteker, 2005; Jacobides & Hitt, 2005). Despite this trend and the surge of interest in capabilities and “competence” (Barney, 1986, 1999; Kogut & Zander, 1992; Winter, 1995; Conner & Prahalad, 1996; Teece, Pisano & Schuen, 1997; Langlois & Foss, 1999; Winter, 2003), to date we do not have a consistent set of tools to expand our intuition beyond the truism that firms specialize in what they are most capable in.

Such a set of tools would be particularly timely, given the surge of interest in changes in the vertical structure in industries and economies: outsourcing and global specialization (Feenstra, 1998), value chain dis-integration and re-integration (Fine, 1998; Christensen, Verlinden & Westerman, 2002; Langlois, 2003; Jacobides, 2005; Jacobides & Winter, 2005) have become more prevalent, and it is becoming clear that both firm heterogeneity and transactional conditions affect scope. Such an effort would help to complement analyses of industry evolution, which have largely neglected the changes in vertical scope (see Langlois, 2003; Winter, 2005). It would also provide an analysis of vertical integration that is not necessarily grounded in market imperfections and oligopoly, which has been the hallmark of the analysis of vertical scope in Industrial Organization (Perry, 1989; Joskow, 2005).

The objective of this paper, therefore, is to provide such a set of tools that expand traditional views on the determinants and dynamics of vertical scope. Drawing on Ricardo’s (1817) theory of comparative advantage, I provide a stylized analysis of how productive capability differentials combine with TC to affect vertical scope at the firm and industry levels. I consider the role of scale, as well as learning curves or the returns on investing in capability development in different parts of the value chain, thus confirming, qualifying and extending previous research (e.g., Stigler, 1951). Using behaviorally plausible formal modeling and simulation, I show that the distribution of productive capabilities between firms along an industry’s value chain determines the vertical scope of individual firms and, by extension, of the industry as a whole; and how such capability differences interact with scale conditions. I then explain how TC catalyze all these factors, and interact with them to determine scope. This model also explains how purely dynamic factors, such as learning curves, endogenous capability improvement, or limits to expansion in different parts
of the value chain interact with TC to determine scope. The model provides a consistent theoretical accounting for empirically observed regularities, such as the use of “mixed governance” (concurrent use of both vertical integration and market procurement), or the stylized fact that industries go through a fairly stable period of integration before specialization emerges.

To achieve these results, I propose a model that departs from existing modes of analysis. Rather than looking at each firm and each transaction in isolation, I consider the firms as a group along with their make vs buy choices, under varying capability conditions, and observe the resulting degree of integration over time. I also allow firms to choose freely between being net buyers or sellers of intermediate goods, or to remain integrated. This shifts from the comparative static, “ceteris paribus” approach, to a systemic analysis of firms that concurrently produce and set prices, quantities and scope in the context of competition.

The paper starts with an overview of the literature on vertical scope, focusing on the growing discussion on how capabilities affect scope; I also briefly consider the literature on how scale affects integration. I outline the theoretical and methodological foundations of the model, and the motivations for its analytical structure. I present the results of the model, and discuss in particular how productive capabilities, scale, learning curves and TC co-determine vertical scope, both statically and dynamically, as well as considering the role of the limits to expansion, and endogenous changes in TC. The paper concludes with a discussion, and with the implications of the model for theory and practice.

**Transaction Costs, Productive Capabilities and Scope: Towards a Systemic Analysis**

How firms determine their vertical scope is a subject that has been of long-standing interest. A branch of microeconomics, and a large part of new institutional economics, has focused on understanding and explaining what affects firms’ decisions to integrate vertically (Perry, 1989; North, 1986; Williamson, 1991).¹

In the last twenty years, much of the existing research on vertical scope is based on Transaction Cost Economics (TCE), which argues that firms choose their scope as a function of the TC they

---

¹ Most of the literature in Industrial Organization (Perry, 1989) focuses on oligopolistic explanations for integration. In imperfectly competitive industries, the pursuit of monopolistic or oligopolistic rents motivates integration: Firms may want to integrate in order to raise rivals’ costs (Salop & Scheffman, 1983), control scarce resources (Porter, 1980), or eliminate multiple marginalization (Salop, 1979; Dixit, 1983), i.e. the problem of setting lower quantities than appropriate for profit maximization through reducing the output twice, once in the intermediate, and once in the final market. Alternatively, integration may be motivated by the desire to gain the ability to price discriminate, or to obtain a strategic upstream supply (Stigler, 1951; Arrow, 1975; Riordan & Sappington, 1987). The analysis in this paper, does not rely on oligopolistic structures and rationales for deciding the firms’ scope, to preserve focus and to show how capability heterogeneity, scale and learning interact with TC in a competitive setting. I thus do not provide an exhaustive review of the IO / oligopolistic motivations; see Perry (1989) for a compendium and Joskow (2005) for an update that confirms the recent predominance of TCE-based researched, reviewed by Klein (2005).
face: if there is a risk of ex post opportunistic renegotiation in dealing with a vendor because a
longterm, asset-specific investment has been made, then that very risk becomes a hindrance to the
potential market transaction (Williamson, 1985; Hart 1995). As a result of the ex ante fear of
expropriation from some future renegotiation, potential supplier firms that are asked to make
asset-specific investments will refrain from such investments. Thus, inasmuch as asset-specific
investments are needed, or inasmuch as specific assets are significantly more productive than
generic investments, the only way to obtain them is through vertical integration. According to
TCE, then, vertical integration is a function of the extent of TC, themselves the result of
underlying asset specificity and hence potential opportunistic renegotiation. Furthermore, relying
on the market introduces frictional costs of locating and monitoring suppliers (Coase, 1937), and
of measuring their outputs, which is more difficult than measuring the output produced in-house
(Barzel, 1982; Masten, 1991).

The focus of research until recently was on demonstrating that TCE provides accurate predictions,
based on the acceptance that asset specificity matters (Shelanski & Klein, 1995; Klein, 2005;
Joskow, 2005). Over the last few years, however, it has become clear that TCE is not a self-
sufficient theory of vertical scope. Williamson, for instance, recommends that the traditional TCE
query, “‘What is the best generic mode (market, hybrid, firm) to organize X?’ be replaced by the
question ‘How should firm A -- which has pre-existing strengths and weaknesses (core
competences and disabilities) -- organize X?’” (1999: 1003). This question was recently also
pursued by Madhok (2002), who suggested that an individual firm’s choice must depend not only
on the characteristics of the transactional conditions, but also on its strategic objectives, the
attributes of its own capabilities, and the governance context it has created.

In terms of theory, though, there is not much to guide our intuition other than that firms will tend
to specialize in areas where they have a comparative advantage (Barney, 1999); while some

---

2 Interestingly, even TCE critics have focused on whether the presumed advantages of integrated governance are due
to opportunism vs other factors, such as better knowledge and information flows, identity, or learning facilitated

3 On the empirical side, the literature on how capability differences drive scope is limited but growing. Walker and
Weber (1984, 1987) found that the most important predictors of sourcing were cost differences between the focal firm
and outside suppliers in producing a specific component. Argyres (1996) provided qualitative evidence on the role of
firm capabilities in integration decisions, observing that in the cable manufacturing business, in addition to TC,
capabilities were a significant driver of vertical scope. The growth of the capability- and resource-based view of the
firm led scholars to consider the role of firm heterogeneity: Poppo and Zenger (1998) considered skill sets and scale
as determinants of outsourcing information technology (IT) services; Leiblein and Miller (2003) found a strong
correlation between insourcing the production of a specific component and past experience producing the component;
capabilities are also seen as a very important driver of the extent of integration in recent research by Hoetker (2005).
Jacobides and Hitt (2005) found that the amount of variance in integration explained by capability differences is at
least of an order of magnitude greater than that explained by TC. Such findings underscore the need for further
theoretical development on the role of capabilities in driving scope.
verbal, appreciative theorizing on how scope depends on capabilities has recently emerged (e.g., Madhok, 2002; Jacobides & Winter, 2005), there are remarkably few concrete analyses in the literature. Yet the current rise in outsourcing and offshoring, and more broadly the increasing vertical dis-integration of production (Feenstra, 1998; Grossman & Helpman, 2002), suggests that a deeper study of how productive capability differences, or issues such as scale or learning curves drive scope is required.

In considering the relative merits of integration and outside procurement, for instance, the nature of Economies of Scale (EOS) is often invoked as the reason for outside procurement being more efficient when viewed at the level of the individual firm. Yet the rationale is not sure to hold when we consider the entire productive system. As Riordan and Williamson (1985: 369) note,

outside procurement might appear to be favored if a firm’s needs for a good or service are not sufficient to support a plant of minimum efficient scale. The same would appear to be true for items that experience economies of scope. But since the firm can always realize the same scale or scope economies as an outside supplier, by selling product that exceeds its own needs on the market, the firm need not and presumably will not experience production cost diseconomies of either kind.

In other words, even if there were EOS in one segment, why should a firm not have the appropriate scale in that segment and in a vertically related one? As Perry (1989) comments, vertical specialization should, at a minimum, require some dis-economy of scale (in one segment) to co-exist with economy of scale in another. So the question of how EOS combine with other factors to drive scope is, perhaps surprisingly, still not satisfactorily addressed.

The analysis of scale and its relationship to integration has been a thorny issue ever since Stigler’s (1951) seminal paper in which Smith’s argument that “specialization is limited by the extent of the market” (note that this refers to aggregate, total demand) is rehearsed. Stigler hypothesizes that as demand grows, there will be enough volume to support specialization. Yet, as Perry (1989: 262) notes, “the difficulty with the Stigler model occurs from trying to capture the notion of specialization. If specialization means only that the production process of the upstream stage is subject to economies of scale, then the vertical equilibrium will degenerate irrespective of the size of the industry.” The problem seems to be that research that followed Stigler’s discussion of scale does not take into account what happens when the entire industry and its dynamics is concerned, as Perry and Groff (1988) and Langlois and Roberts (1995) pointedly suggest. As Dufeu (2004: 4

A full discussion of Stigler’s hypothesis beyond the scope of this paper. For the interested reader, we recommend Perry and Groff (1988) or Dufeu (2004). The Stiglerian argument is that if specialization requires higher fixed costs, but leads to smaller marginal costs, then market growth will allow more efficient and also more specialized production to occur; a small market would be associated with integration. The evolutionary logic of this argument is that as an industry grows, stages with increasing returns will be spun off, leading to specialization as a function of the increase in total demand. This argument, which spurred some empirical literature on market size and specialization, has proven fairly elusive, not only empirically, but also analytically. White (1978) and Perry (1989) suggested that
4) puts it: “Perry and Groff (1988), taking into account White’s (1978) criticism of Stigler’s model…show that… the dynamics of vertical integration cannot derive exclusively from the existence of substantial economies of scale” (emphasis in original). Summing up, the arguments made about the link between EOS and integration seem not to hold when we consider the entire productive system.

Considering the entire productive system becomes even more important when we weigh transactional conditions. As Grossman and Helpman (2002: 86) recently suggested,

Economists who have studied the make-or-buy decision have focused on the bilateral relationship between a single producer and a potential supplier. Beginning with Williamson [1975, 1985] and Grossman and Hart [1986], a body of literature has developed that clarifies the role of transactions costs, asset specificity, and incomplete contracts in guiding a firm’s choice of whether to undertake an activity in-house or to seek to fill the need from the outside. Yet, influential as this work has been, it can provide only a starting point for understanding cross-sectional or cross-regional differences in outsourcing behavior, or for evaluating explanations for recent trends. This is because a decision-theoretic approach treats the industry environment as given, and thereby neglects the interdependence among the choices facing the various firms in an industry. For example, the attractiveness of outsourcing to a certain producer may well depend on how many firms potentially can provide the inputs it needs, which in turn may depend on whether other firms in the industry have chosen to be vertically integrated or to buy their inputs from others.

This observation raises a simple yet fundamental concern: in an economy, or an industry, the “system” (the sum total of firms and their decisions) must “add up”. That is, every purchase (of intermediate product from the market) must also be a sale (of intermediate product by some other firm). Obvious as this may seem, much of the existing research, focused as it was on establishing the relative importance of, say, transactional conditions, has neglected this point.

Consider, for instance, Riordan and Williamson (1985). Their analysis (re-iterated in Williamson, 1996: 65-74) starts with the examination of how asset specificity affects TC. These authors go on to argue that in addition to considering TC, a firm should also consider the differences between its own production costs, and the production costs of “the market”, the former always being lower. In their words, “Production cost differ between firm and market, whereby the firm experiences a production cost penalty when asset specificity is slight but these cost differences asymptotically

---

taking the entire industry, the logic in the Stiglerian argument is hard to establish without additional, rather ad hoc assumptions being made. As Perry (1989: 232) notes, “As soon as demand has grown to the point that two downstream firms are more efficient than one, the upstream stage would be spun off. Vertical disintegration would prevail thereafter”. Notably, most attempts to support Stigler’s (1951) hypothesis have been founded on ologopolistic structures with differences between fixed and variable cost conditions of integrated vs specialized firms. The trade-off between integration vs specialization in such models is between the technological benefits of specialization and the foregone benefits specialization implies as a function of multiple marginalization; aggregate demand growth (and entry conditions) affect either or both aspects, and as such are shown to affect integration and the vertical equilibrium. See Elberfeld (2002) or Dufeu (2004) for illustrations.
approach zero as asset specificity becomes great” (Riordan & Williamson 1985: 369). This seemingly innocuous assumption, though, has substantial implications. Essentially Riordan and Williamson imply that any other firm (i.e., the one behind the veil of “the market”) is never worse than the focal firm in terms of production costs (recall that TC are added to production costs in Riordan and Williamson’s formulation). So this reasonable device of juxtaposing “firms” with “market”, chosen for analytical convenience, and pervasive in the literature, leads to a non sequitur when seen from the vantage point of the entire industrial system.

The point is worth stressing. In any industry, in any universe of trading partners, it can never be that everyone is worse than everyone else, becoming almost as good at the asymptote. It must be that some focal firms (those that will in equilibrium be outsourcers) will necessarily need to be superior to other focal firms (those that in equilibrium will be outsourcees) in terms of production costs, for there to be any trade. After all, the “market” does not produce anything; behind the thin veil of the “market” is another firm, whose very decision to specialize or not hinges on competitive dynamics. Thus, to take the theory forward, we must consider the systemic interdependencies between TC and other factors. The emphasis on considering the entire industry, then, brings us back to the need to construct a model that “adds up”, where scope is endogenously derived along with the other key production decisions. Such an approach almost inescapably means we have to abandon closed-form solutions and consider computational techniques; yet the sacrifice in analytical elegance could handsomely pay off if we constructed a consistent model.

In building a model of the sum total of producers, we should also include the role of heterogeneity of different firms in terms of production costs, an issue that has been relatively neglected, as Richardson (1972) and Langlois & Foss (1999) argue. In Demsetz’s words, “The decision [to produce] hinges on the internal costs of production that burden the potential purchaser and supplier… The emphasis that has been given to transaction cost… dims our view of the full picture by implicitly assuming that all firms can produce goods or services equally well.”

---

5 Grossman and Helpman (2002) provide a very different model from mine. Their one-period formulation considers a universe of firms that could decide to be integrated or specialized. Grossman and Helpman derive the equilibrium vertical scope in an industry, arguing (a) that vertically specialized firms always have lower fixed and marginal costs than integrated firms in each part of the production process (that is, that in terms of productive capabilities two vertically co-specialized firms are always superior to an integrated firm); but on the other hand, that (b) the integrated firms need to find “matches”, which is an uncertain process, with firms deciding, in a one-period game, where to participate (in the integrated or specialized pool); and (c) that once specialized, they have no or only a limited “after-market” if they do not find a perfect match. On the basis of individual firms’ expectations of integration or specialization maximizing their profits, Grossman and Helpman predict when we should observe many or few choosing (in a one-period game) to specialize or integrate. My model is fundamentally different from theirs in providing a broader set of potential conditions; in suggesting that firms do not need to commit to whether they will be specialized or not; in considering that search and matching difficulties are only one of the possible issues at hand; in looking at the differential impact of such search costs conditional on other differences (scale, capability endowments etc); in allowing endogenous pricing; in considering different potential capability and transactional structures, as they interact to drive scope; and in incorporating dynamics and learning, and showing the selection mechanism at play.
Winter (1988) concurs, suggesting that firm scope, like any other business activity, is related to the particular way a firm “does things.” He observes that “firms are repositories of productive knowledge…that involve[s] idiosyncratic features that distinguish [them] even from superficially similar firms in the same line(s) of business.” (Winter, 1988: 175)

Coase also raises concern about the lack of attention to the heterogeneity of production costs: “[This] has tended to submerge what to me is the key idea in “The Nature of the Firm”: The comparison of the costs of coordinating the activities of factors of production within the firm with … operations taken within some other firm.” (1988: 38, also quoted in Williamson, 1999).

So not only do we need to consider the entire system, we also need to take into account firm heterogeneity. The existence of such heterogeneity should be obvious to management scholars, who would argue that firms would tend to specialize where they have comparative strengths, it would be hard to gainsay this. But, the question that emerges, is can we move beyond this truism? And how can we integrate scale, and the possibility of learning curves and TC, all of which we know to be important?

Drawing an analogy with international trade theory (Ricardo, 1817; Deardorff, 1980), I would argue that to understand patterns of specialization we have to examine the comparative advantage of all those who do participate or could participate in the production process, and to also consider issues of scale, as well as learning curves or capability development. Another important reason that systemic analysis is called for is that it enables us to consider the implications of competition and selection for vertical scope. Vertical scope is only one of the choice variables for a firm, and our analytical framework must allow for this; in other words, rather than presuming the existence of a given number of buyers and suppliers in an intermediate market, we must allow for firms choosing whether, and how much, specialization is appropriate for them, which segment(s) they will focus on, thereby endogenously and incidentally determining vertical scope, as a function of both TC and capability differences. Furthermore, even if the objective is to understand an individual firm’s decision in terms of vertical scope (i.e., “do I specialize or not?”), we need to understand what shapes the firms’ environment and provides the firm with the cost-benefit calculus of specialization (“specialization depends on my relative productive capability positions which depend on what others in the industry can offer; and that, in turn, depends on competitive dynamics”). To achieve this, I propose a model and attendant simulation to explore the following questions:

1. How do differential capabilities and their distribution affect the choice of vertical scope?
2. How do scale conditions in different parts of the value chain interact with capability differences to drive scope?
3. What are the roles of learning curves, of capability improvement structures and limits to expansion in shaping integration?

4. How do TC combine with capability distribution, scale, learning curves and capability development to determine integration?

The model builds conceptually on dynamic programming and activity analysis, using a Mixed Complementarity approach (Ferris & Kanzow, 2002). It has a Marshallian structure, in that the factors in the short term that are fixed are altered by firms in the long term. In keeping with the “history-friendly” approach (Nelson & Winter, 1992; Malerba et al., 1999) particular attention has been paid to the model’s behavioral plausibility, within the bounds any formal analysis inescapably imposes. The next section provides the intuition behind and verbal exposition of the model, focusing on its motivation and structure; the full analytical layout and mathematical formulations are included in an Appendix at the end of the paper.

The Model: Structure and Mechanics

Model Overview and Motivation. The aim is to explore what happens in an industry where firms have differing productive capabilities in different parts of the value chain, and then to consider how such capabilities combine with differences in scale, learning curves and TC to drive scope. To do this, I constructed a model with two vertical segments: an “intermediate good/upstream” segment, and a “final good/downstream” segment. (Note that the terms “intermediate” and “upstream,” as well as “final” and “downstream,” will be used interchangeably in the remainder of this paper.) Figure 1a summarizes the basic structure.

Within this context, each firm has a priori the choice of producing in either or both of these segments. Intermediate production is a necessary input to final production in fixed proportions. Firms select whether and what they produce, and whether they consume their intermediate product, sell it to other firms, or buy it from other firms, thereby determining their vertical scope. They also decide the prices at which they trade, both on the intermediate and the final markets. In order for a firm to produce in either segment, it needs resources -- a segment-specific bundle, for which it pays a price. The supply function for both upstream and downstream resources is common to all participants, who compete for resources and pay the market price. In my simulations, I also consider the impact of changing some key parameters of the model - e.g. the relative EOS in the two segments, upstream and downstream. While firms can choose their

6 In terms of prior modeling literature, a notable yet little cited contribution is Rubin (1973). Rubin is interested in formalizing Penrose’s (1959) insights on limits to expansion and applying them to mergers and acquisitions (M&A) and corporate control, which has some similarities with the model presented in the next section. Both models use programming analysis, treat resources and capabilities as distinct elements, and posit that limits to expansion matter.
level/scale of production endogenously, I explore how changing the technological economics of production affects integration.

Crucially, firms are endowed with heterogeneous productive capabilities, both in terms of upstream/intermediate and downstream/final production. Given the relative novelty in expressly considering capabilities in driving scope, a short theoretical aside is called for. Productive capabilities are defined in the model as the way in which (i.e., the efficiency with which) resources are turned into products – something akin to what Winter (2003) recently dubbed “zero-level” capabilities. Such productive capability differences, largely ignored in the literature, are quite important in practice (cf. Lieberman & Dawhan, 2001). Productive capabilities rest on the firm’s general and specific knowledge of how to do things (Richardson, 1972; Teece, Pisano & Shuen, 1997) as well as the specific investments and complementary assets (equipment, training, retention of key personnel, etc.) required to put that knowledge to work (Barney, 1986, Winter, 1995). In this perspective, heterogeneous capabilities can arise as a result of a path-dependent learning process, in which there is abundant opportunity for various contingencies to shape the way of doing things that ultimately emerges (Winter, 1988, Levinthal, 1997), even if all firms have access to homogeneous resources – as they do in the model.

I also consider endogenous capability development rules, and how they affect scope. First, I will allow firms to develop their capabilities, and improve them either as a result of learning curves (Argote et al., 1990), where an increase in cumulative production leads to a dynamic improvement in capabilities, such that production begets capability; I consider how scope changes when learning curves exist in one or both of the vertical segments. I then consider how my results change as a result of another endogenous capability development mechanism whereby profits (from the most effective firms) can be re-invested to improve capabilities, and once again I consider what happens when capabilities both up- and downstream improve, or when they improve only in one segment – upstream.

The structure of the model consists of the short run – which I dub the “static model”– and the long run, i.e., the “dynamic update.” In the short run, following the Marshallian tradition (1926 [1890]), the factors of production are fixed: firms have a “maximum endowment” of the resources they can use – a limit on their capacity (a maximum number of branches, a limit on number of employees, etc.). Each firm knows its capabilities and TC, and decides on production of intermediate and final goods. A firm, of course, may elect not to fully utilize its capacity, if it is

---

7 Note that an advantage in terms of productive capability leads either to superior quality, or to inferior costs per output of unit. A difference in quality can be modeled as a difference in cost through quality adjustment. Specifically, we can consider the firms’ relative productive capabilities in terms of their costs for a quality-adjusted unit of output.
not profitable, as indicated in Figure 1b. The amount of planned production will then determine the prices of resources (through aggregating firms’ demand and meeting a price elastic supply); the price of the final good (through aggregating firms’ supply schedules and meeting an elastic demand); and the price for the intermediate good, if any trade happens (as a result of firms’ choice in participating in the intermediate market as net buyers or sellers). The system will equilibrate when supply meets demand in all markets (upstream and downstream resources; final and intermediate good market, whether trade happens or not), and that will determine production.

As an illustration, if this were the cement industry (cf. Ghemawat & Thomas, 2003), the model’s upstream sector would be cement production, for which a resource bundle is needed (klinker, labour, energy, capital). The upstream capability of a firm would be the efficiency, i.e. the cost of converting this resource bundle (which can be bought on the market, and whose price is a function of the resource bundle’s industry-wide demand) into cement. The downstream segment would be the cement mixing business, for which cement is needed in fixed proportions, and for which another, additional bundle of resources (inert materials, energy, labour, capital) is needed to produce pre-mixed cement. The downstream capability would be the efficiency with which the downstream resource bundle (whose price is endogenous to industry-wide demand) and the upstream good (cement) are turned into pre-mixed cement. The price of pre-mixed cement would depend on supply, which is a function of what the industry participants can offer, and demand from the final customers, who are price-sensitive. Firms can trade in the intermediate good (cement) by selling or buying it in any quantities, at prices that are endogenously set by the firms who want to buy or sell cement. Firms also have in each period a maximum capacity endowment (number of plants upstream and mixing facilities downstream which they can operate). However, within that maximum endowment, they are free to choose their scale, taking into account the prevailing EOS.

The next part of the model is the dynamic update, where the firm plans its capacity for the next period. Each firm considers changes in its “fixed factors,” i.e. its maximum resource endowment, and decides it’s the pattern of expansion or contraction per segment. Its production or its profitability will also potentially affect its capabilities in the next period. Learning curve effects or profit re-investment in capabilities, which may differ between segments, will affect each firms’ capability in the next period. With these new Marshallian “fixed factors” in place, the next period begins, competition and production happen, and the cycle repeats itself.

Continuing our cement example, the dynamic part of the model would be akin to the adjustment in firms’ capacities, both in the upstream and the downstream segment. On the basis of how
successful they have been, firms would plan their expansion. Also, on the basis of their production and profit in the previous year, they may change / update their capabilities to incorporate the learning that has taken place in either vertical segment, as well as account for any re-investment in capability development. This provides the new capacity limit (plants, maximum number of employees, etc.) and a new set of productive capabilities for both the up- and down-stream, with which each firm competes in the next period.

Still using our industry example, the key question that the model would investigate is whether firms would specialize in cement production, or in mixing, or would integrate by being active in both, as a function of the distribution of their productive capabilities (efficiency of turning inputs into outputs in cement production and cement mixing) and of the level of TC in each period. Also, it can show how such specialization in cement production or mixing evolves over time, as a function of both individual firm’s choices related to scope, and changes in the structure of the industry as a function of competition and natural selection (e.g. if specialized firms grow more than integrated ones, this will push the industry towards greater specialization).

While the full model is illustrated in the Appendix, I provide here a description of the analytical structure, starting from the individual firm’s decisions in each period, moving on to how firms interact through the final and intermediate markets, and concluding with the “dynamic update” part of the model.

Static Model, and the Mixed Complementarity Solution

In the static structure, each firm’s objective is to jointly maximize the profits of both its intermediate and final goods divisions, by deciding the structure of its production. It can produce intermediate and/or final goods in any combination: To produce final goods, of course, it needs a given quantity of intermediate goods. It can use its own intermediate goods, buy from the market, or both; similarly, it can decide to produce more intermediate goods than it needs, and thus become a net seller to the market for intermediate goods. The firm has the option of participating in the market for intermediate goods, either as a seller or as a buyer; the market for intermediate goods then either clears at a price, if there is trade, with firms either selling or buying from a central trade pool; or if there is no price that can make both buyers and sellers participate, no trade

---

8 The reader familiar with the economic literature will recognize that this structure is a generalized version of activity analysis, usually examined through linear programming. The examination of the firm as a bundle of activities, each requiring resources, has a respectable pedigree (see Dorfman et al., 1957, chs 6 and 7 for seminal expositions), even though not much has been written recently on these topics. Furthermore, the connection between the capability- and resource-based views that have gained prominence in the field of strategic management, and this set of existing tools and analyses has not yet been made.
occurs. Last, the final good is not differentiated and the total supply in that market is determined by aggregation of the supply of the firms in the industry, whereas total demand is determined by an elastic demand function.\(^9\)

The cost of the resources the firm uses in both segments is endogenously determined through the market for resources, where the demand for resources is the aggregate of all the individual firm’s demands, and the supply is elastic to price. In mathematical terms, the static model consists of a set of non-linear optimization problems (firms jointly maximize profits in the upstream and downstream divisions). Therein, firms select how much they produce up- and down-stream, as well as whether they buy or sell intermediate goods, and at what price, under given constraints. In addition, the global equilibrium conditions ensure that the intermediate and final goods and resources markets, which link all firms together, can clear. The resulting problem belongs to a class of models known as Mixed Complementarity Problems (MCP), for which solution algorithms exist in the computational general equilibrium literature (see e.g., Ferris et al., 1992 or Capros et al., 1998 for applications; for additional details, please consult the Appendix).

To recapitulate, firms start each period with given industry-wide TC, maximum capacities, and capability levels. Solving their profit-maximization for both the upstream and the downstream divisions, they decide on the level of production of both the final and the intermediate good, thus affecting its degree of integration (vs use of the market).

**Dynamic Update: Expansion of Endowment, and Capacity and Capability Evolution**

So far, I have described what happens within each period; in the remainder of this section, I describe the “dynamic update” part of the model. At the end of a period, firms change the factors that were fixed. Specifically, they first update their resource endowments. If they have used up more than a certain percentage of their endowments, they expand; if they produced below capacity, they shrink to reach a level closer to actual demand. If a firm expects that its optimal production quantity in the next period will be roughly the same, it will follow that the firm will require its future maximum capacity be somewhat higher than current capacity, which we denote as the “capacity buffer”, which means that each firm will want to have a maximum capacity that is a multiple of (i.e., percentage over) the capacity used in the previous period.

This expansion structure was inspired by discussions in the field, and by the perception that all firms set targets of relative growth, planned to help them meet higher demand in the next period.

---

\(^9\) We can model differentiated products or spatial competition by introducing stickiness, but this would greatly complicate the model and make the initial conditions much more important. Similarly, varying the elasticity parameter provides some interesting results, but exploring them would make the exposition overly complicated. As such, analysis of elasticity and differentiation has been left as an extension to this work.
or to win market share from their competitors. Even in a declining demand scenario, firms still do not fully adjust to the downsides of demand, and this formulation provides a simple heuristic in terms of expansion policies. (Whether that extra capacity will be used next period or not will of course be determined during by applying the “static” model.) The “growth aim” of firms (i.e., the value of this capacity buffer) is a parameter that is not varied in the model, since it does not change the qualitative results of the model (its value was set to 1.1); its main impact is to “speed up” industry evolution, as more aggressive expansion allows the effective firms to dominate more quickly and as such can be thought of as the intensity of the selection environment.  

The next change is the endogenous change in capabilities. Firms’ capabilities can change through cumulative production, via a learning curve formulation where the logarithm of the volume produced enhances their skills (Argote et al, 1990); in the model, I consider how the existence of learning curves in one or both of the vertical segments affects specialization. Second, capabilities can improve through firms re-investing their profits in capability development. The rule of capability re-investment that I set first apportions the profits to be re-invested according to the share of production in the previous period. Each sector receives proportionate funding for capability development, and I posit that such monies serve partly as funding for research and development (R&D) (in that the capability improvement is partly a function of absolute profits invested) and partly as funding for training (in that capability improvement is also a function of the profits invested per units of capacity). I then consider how integration changes when we introduce such endogenous rules in either both or only one of the vertical segments (the other, e.g., not yielding return to investment).  

The discussion of expansion brings us to the analysis of entry and exit. Exit, in this model, happens when firms do not use their productive capacity, and gradually or even abruptly withdraw from the market and wither. However, any firm that “started” in the industry can re-enter. Also, “entry” can occur through an increase in the productive capacity of the existing firms. Of course, in addition to “capacity entry”, we could have de novo entry of firms that come into the sector. This is an option I did experiment with, and on which I did a number of robustness checks. Entry ends up not affecting any of the qualitative results of the paper inasmuch as the entering cohort, at any point in time, has the same structure of capabilities (in terms of upstream / downstream correlation) as the initial cohort. The only situation where entry might lead to some discernible differences (which still preserve fully the qualitative intuition) is in the case of asymmetric EOS; as such I explicitly consider the role of entry in this scenario in the results section. Given that entry introduces a degree of modelling arbitrariness (e.g., should entry be accompanied by the demand growth that will absorb the extra capacity that is added to the industry? How large should entry be compared to incumbents? What should motivate it?), and some bumpiness in the graphs (due to the fact that new firms are added, temporarily affecting some of the competitive dynamics), I opted not to include it, since it does not affect our results.

I also experimented with other re-investment rules, such as apportioning the profits on the basis of firms’ comparative advantage, or, alternatively, investing to redress firms’ comparative weakness. Two key observations are worth including here. First, the basic insight on the impact of capability re-investment rules on integration is preserved, whatever their specification. Second, there is increasing arbitrariness in more “strategic” rules, and their discussion might best be left as an extension to the paper. Third, the impact of any re-investment rules is fairly intuitive when compared to what I propose as a benchmark. For instance, re-investment which “primes” the area of advantage of a firm will lead ceteris paribus to an increase in vertical specialization; whereas re-investment, which
This completes the discussion of capacity and capability updates, given plausible bounded rationality conditions. With the new fixed factors in place, the next period is run; output, prices, profitability and integration are determined; firms take these results into account in deciding their expansion for the future and the cycle can be run for an indefinite number of periods.

**Modeling Transaction Costs.**

In the description of the model, I briefly mentioned that TC are conceptualized as the net additional costs that are imposed when using the market. This was motivated both by modeling parsimony, and by a desire for consistency with various strands of literature that analyze TC. Williamson (1985, 1991, 1999), for instance considers that the choice of governance structure hangs in the balance between the costs of bureaucracy and muted incentives $B$, which are inimical to integration (Williamson, 1985: Ch 6), on the one hand; and the potential costs $TC$ of using the market that create the risk of re-negotiation and limited adaptation, on the other. The contribution made by Williamson was to recognize that these market-based TC themselves are the function of asset specificity, so that $TC$ increase as asset specificity increases. As bureaucratic costs were scantily studied by Williamson and his followers (see Zenger & Hesterly, 1997 for an extension), the focus had largely been on how decisions to integrate or specialize hinge on the changes in asset specificity, themselves a function of the production technology (and its attendant need for relationship-specific co-specialization) and of the existence of appropriate contractual technology (Williamson, 1999; Madhok, 2002). This suggests that we can construct a measure of the “Net Costs” of using the market, which are themselves the difference between the actual TC (in essence, the net present value of the risk of expropriation, or the estimated costs of mal-adaptation, as Williamson suggests), and the Bureaucratic costs $B$, so that $NC = TC - B$. Indeed, as Williamson makes clear in a diagram which is repeated in many of his seminal contributions (e.g. 1985: 91; 1996: 69), with low asset specificity $TC < B$; whereas with higher asset specificity, $TC > B$ which leads to the adoption of increasingly “integrated” governance interfaces, which have higher $B$ but lower $TC$. This approach of considering the “net costs” of using the market is 12

12 In the model, firms react rationally to their results within-period; the model’s structure does not make any heroic assumption about their abilities to decide production and resource use, largely as several factors are fixed in the short run. Firms also respond rationally to their expansion problem, but in a myopic way. They do not strategize with regard to resource markets and their evolution; they simply try to expand as much as they think will be profitable. A final (and major) element of inter-temporal myopia is the fact that in deciding their expansion and hence their capability evolution, firms implicitly assume that prices will remain similar to what they were in the previous period. They also do not strategize in order to change the market structure; they react rationally, but are not omnipotent strategic actors. So in this version of the model, the inter-temporal optimization is consistent with a boundedly rational set of actors, who are proficient in their “local” economic and strategic problems, and adept (but limited and not perfect) in their inter-temporal decisions. An extension is to consider greater foresight and strategic behavior.
also advocated by Gibbons (1999).\footnote{As both Williamson (e.g. 1985: 33; 1996: 63) and the incomplete contracting and organizational economists argue (see Hart, 1995), a firm \textit{can} have the choice of a more generic input or output configuration, which has lower value / utility but lower asset-specificity (use of generic dyes in automobile manufacturing, for instance); or it can use a more specific technology with a higher value. The challenge is that for the “higher value”, asset-specific technology to obtain, integration may be necessary for two reasons. First, the purchasing firm will integrate to protect itself from mal-adaptation in the future. Second, the potential supplier might never find it appropriate to make the necessary investment to produce a co-specialized good, for fear of expropriation by the buyer in the future. This second, complementary rationale, has been emphasized by incomplete contract theorists (see Klein, Crawford & Alchian, 1978; Grossman & Hart, 1986). In my formulation, I model the lack of productivity that results from lack of co-specialization as a cost, i.e. a part of the TC itself. I then consider what happens when we change the level of TC associated with market transactions that are imposed to an appropriately customized solution.}

What I posit in this paper is that we can consider worlds with greater or lower asset-specificity and greater or lower contractual completeness, either of which will lead to a higher “net cost of using the market”. Using a “reduced form”, I am agnostic on what \textit{causes} the different levels of TC. What I do in the model, is to look at what happens when we \textit{change} these contractual conditions – when we increase or decrease the net costs of using the market, whether this change is due to standardization of technology (which leads to a reduction of asset specificity or risks of hold-up) or to improvement in the institutional background (which makes contracting through the market less risky). I thus vary the overall level of TC, which, as Williamson (1996) and Grossman & Hart (1986) have suggested, are a net outflow from the system; they represent the cost of “maladaptation”, including the important “risk of expropriation”, which can be expressed in Net Present Value terms. Such “Net Transaction Costs”, then, can be \textit{registered} as a “net tax”, itself a function of asset-specificity: A high “Net TC” will mean that the difference between what you can get through the market (as a function of imperfect co-specialization or latent risk) and what you can achieve in-house, production costs aside, is great.

The conceptualization of the costs of using the market as a “Net Tax” has the added advantage of being consistent with other economics-based research on governance: Such a net tax could also encompass the frictional Coasean costs associated with using the market (such as setting up purchasing divisions, searching the market, etc.); or the extent to which the measurement costs between two stages of the production process are higher when it is performed outside the boundaries of one firm (Barzel, 1982; Masten, 1991), thus leading to reduced efficiencies in vertical specialization.\footnote{Note that this is a “net tax”- that is, a tax that represents the \textit{additional} risks of market transaction when compared to coordinating the two parts of the production process via bureaucracy, i.e., common ownership (see, e.g., Gibbons, 1999). In its purest form, the TCE argument could be that such a tax, for perfect markets, can even be \textit{negative}- that is, in-house production can be less effective (regardless of productive capabilities) as it is riddled with bureaucracy, politics, and inefficiencies that markets do not have (Williamson, 1985: Ch 6, Milgrom & Roberts 1992). I thus “normalized” TC to start at zero, both for expository ease, and because a negative tax would be a perverse element in such an economic model: it would mean that trading an intermediate good in and of itself introduces a source of value to the system, and the model would collapse in a world of infinite trading to reap the returns of the negative tax.}
The use of such a reduced-form “net tax”, which varies from zero to a high percentage of the value of the intermediate goods, is also consistent with non-TCE approaches. For instance, part of the “tax” may reflect the fact that the market does not afford an economical set of communication, coordination, and language codes (Arrow, 1974; Pelikan, 1969), or timely reaction (Langlois, 1992), and as such managing across firm boundaries is costlier than managing within them. Such costs can be offset by the expected gains from trade due to production cost differences; indeed, it is this tradeoff between TC and productive capability differences that is at the heart of this model.

I should, however, identify and justify a simplifying assumption that I make: I do not vary this “net tax” by firm. While in principle I could do so, I choose to consider in this model the impacts of industry-wide variation in transactional conditions, so as to understand how changes in TC that affect all firms interact with other factors (such as scale, learning curves, etc.) to drive integration. Of course, in the real world not only do we observe differences in industry-wide means (i.e. settings in which it is easy to transact and ones in which it is not); but also variations in the comparative ability of firms, or pairs of firms to transact with each other (see Dyer, 1996). While modeling such firm-specific, or dyad- or network-specific effects is possible, in order to preserve focus I do not do it in this paper. I briefly show how my results would be qualified by such heterogeneity in the discussion section.

Finally, I consider some additional endogenous features of TC changes as extensions to my model. First, to address the concerns of the “small numbers bargaining” problem, which may add challenges over and above the other transactional difficulties imposed, e.g. through asset specificity. Thus, I consider the challenges posed by the endogenously determined potential number of buyers and sellers by using a “dissimilarity index” of concentration that compounds the “base Net TC”. So when there is a substantial imbalance (small number of sellers selling to many buyers), the “final net TC” increases, to capture some of the additional pressures posed by such potential market transactions. Alternatively, to accommodate improvements in TC conditions as a function of experience in contracting (see Mayer & Argyres, 2004), I use a “learning curve” in contracting, whereby increased trade endogenously reduces the levels of TC.

Research Design: Research Questions, and Resulting Variables and Scenarios.

The model allows us to consider a number of different issues, such as how profits or resource prices are affected by capability distributions and TC, both statically and dynamically. The focus in this paper, though, is narrower: it is on what determines vertical scope at the firm and the industry level, and how that changes over time. The objective, therefore, is to understand how different productive capability distributions, and different conditions in terms of scale, learning
curves, investment in capability rules, and expansion limits along the value chain, interact with TC to determine the scope of individual firms, and, as a result, their degree of vertical integration/specialization at the industry level. More concretely, I consider how changes in TC affect the industry’s vertical specialization, and then study how different capability distribution, scale, learning curves and expansion limits, structure, moderate or mediate these results. Since the objective is to understand how scope evolves over time, the analyses consider the evolution of vertical scope in the industry as their dependent variable. Specifically, in the three-dimensional graphs (Figures 2a to 9), which illustrate the model’s results, one axis ($z$) is vertical specialization at the industry level, anther axis ($y$) is time, showing how specialization evolves over 30 periods; and the third axis ($x$) represents TC, which vary from 0 to 100% of the cost of intermediate products, in ten 10% steps. Thus, each figure describes vertical specialization over time (for 40 periods): for a range of TC values, and for a given distribution of productive capabilities, scale, learning curve conditions, etc.

Furthermore, each figure corresponds to a particular set of capability distribution and expansion costs, which I refer to as a “scenario”. I start with a “core” of three scenarios (1a-1c, depicted in Figures 2a-2c), each responding to a particular set of initial capability distributions (high / zero / negative correlation between individual firms’ upstream and downstream productive capabilities at $T=0$). In this set of scenarios, which is the simplest one, I set EOS in the two segments to be equal; there are no learning curves, either up- or downstream; there is no endogenous capability development process; and the expansion limits are modest and identical in the two segments.

I next go on to vary some of the variables that are of theoretical interest. The next set of scenarios (2a-2c), depicted in Figures 3a-3c, consider the same “base set” of scenarios, but with the difference that EOS in the two segments are varied: I consider higher EOS upstream (in manufacturing) than downstream (in distribution). The following set of scenarios (3a-3c, in Figures 4a-4c) is based on the original scale conditions, but introduces another “treatment effect”: I now examine how our three core scenarios are affected when there are learning curves in both the upstream and the downstream segments. Scenarios 4a-4c, in Figures 5a-5c, include learning

15 Specifically, all the figures consider the ratio of intermediate good traded through the market over the total production of intermediate good, be it produced in-house or procured outside, which of course is the weighted average of vertical specialization for all firms in the industry.

16 Strictly speaking, the initial capability distributions I examine in this model are exogenous. However, this very structure of the capability distributions depends on some other factors, such as the knowledge base. Thus, I would argue that if two vertically related segments rest on the same knowledge base (e.g. chemical engineering), then we should expect a strong correlation between upstream and downstream capabilities; whereas if the two segments rely on different knowledge bases (e.g. chemical engineering and distribution/sales prowess) then it is more likely that the capabilities will not be similar up and downstream. Thus, my research suggests that we should study what determines the pattern of capability correlations along the value chain. In this model, I only explore the consequences of permutations of these capability distributions; but we have much to learn by understanding what brings these about.
curves in only one of the two segments, e.g. learning curves only upstream. The fifth set of scenarios (5a-5c, in Figures 6a-6c) illustrate what happens when endogenous capability development rules applying to both segments are introduced; while the sixth set (6a-6c, Figures 7a-7c) shows what happens when capability improvements occur in only one segment (upstream). Using these sets of scenarios therefore allows investigation of how scope changes over time with the introduction of endogenous changes in TC, either through learning, or through concentration and increases in TC as a result of a reduction in the number of trading partners. I also show how introduction of a “focus benefit”, which increases a firm’s productive capability if it focuses on one segment, affects scope, although the respective figures are omitted to save space.

Finally, I consider some further interesting cases. First, I introduce asymmetric expansion costs, which allow one segment to grow more easily and more quickly than the other (Figure 8). Second, I introduce a much higher level of initial demand to see how this affects the evolution of industry scope (Figure 9); these scenarios relate to theoretically interesting empirical regularities.

All the scenarios have the same overall mean and initial variation in productive capabilities by segment; that is, if we consider the variance of productive capabilities between firms in each segment (separately upstream and downstream) in the initial period, the variance is the same for all scenarios and is identical for both upstream and downstream. However, the scenarios differ with regard to the correlation between individual firms’ productive capabilities in the up- vs the downstream segment. In the “asymmetric capability distribution” scenarios (1a, 2a, 3a, etc.), there is a negative correlation between the upstream and the downstream initial capability of the firms in the industry such that a firm is good either upstream or downstream. Conversely, in the symmetric capability distribution scenarios (1c, 2c, 3c, etc.), there is a very high positive correlation between the initial productive capabilities of the two vertical segments – in other words, whichever firms are good upstream are also good downstream. In the zero-correlation distribution scenarios (scenarios 1b, 2b, 3b, etc.), there is no systematic pattern of efficiency up-vs downstream; some firms perform well in both segments, others in only one or the other – in other words, being good upstream is not correlated with being good downstream.17

Furthermore, the differences between the sets of scenarios (1a-1c vs 2a-2c) rests on how other factors, which can be thought of as “treatment effects” (industry-wide conditions in terms of prevailing EOS, learning curves, endogenous capability development, TC endogeneity rules, etc.) affect scope. Through this design, we can jointly examine how the capability structure interacts

17 The fact that in the all scenarios we have the same overall aggregate capability variation in the industry allows us to hone in on the particular implications of correlation structure between productive capabilities along the value chain.
with transaction costs to affect integration; and consider how our “treatment effects” (scale, learning, re-investment) mediate this relationship. Tables 1, 2 and 3 summarize the research design.

Include Tables 1, 2 and 3 about here

**Parametric Choices and Robustness Checks.**

Substantial effort was made to verify that the model’s results hold for a large part of the solution space, and that they are robust to changes in the parameters of the model. To achieve this, I considered how the results of the base model (scenarios 1a-1c) would change in qualitative terms if I varied the parameters within reasonable ranges. As Table 3 suggests, none of the results was reversed through the wide ranges considered; additionally, how a parameter affected the results, was economically interpretable. Table 3 summarizes parameters, ranges and results.

Specifically, the elasticities of demand for the final product and of supply for resources, intermediate and final, were set to \( \varepsilon_F = \varepsilon_{SI} = \varepsilon_{SF} = 1 \). Elasticity of production was set, both for intermediate and final goods, to \( b_I(i) = b_F(i) = 0.9 \).\(^{18}\) To construct the “strongly negatively correlated” scenarios, I set to the correlation coefficient between the upstream and downstream segment capabilities as \( r = -0.99 \); \( r = 0.99 \) for the “strongly positively correlated” scenario; and \( r = 0 \) for the uncorrelated scenario. The scenarios reported here represent different “runs” of the model, for 21 firms over 30 periods. A range of parametric values (reported in Table 2) was considered for the more than 67 different versions of the base model, and the results discussed in the next section still held; that is, the qualitative results reported were found to be robust to parameter permutations.

Finally, in all the scenarios, in the first period I set all firms in the model to have balanced resource endowments (i.e., identical resource endowments up- and downstream), and as such the industry was initially predisposed to vertical integration. The model then considered whether integration would be maintained, or if and when specialization would emerge. This modeling structure is motivated by Stigler’s (1951) description of the evolution of vertical integration, where the industry starts out being vertically integrated and then, as opportunities arise, it dis-

\(^{18}\) Note that the elasticity of production is set to be identical across firms, as it represents “common technology.” We tried a range of values (0.5-1.2) which did not materially affect the results. More specifically, an elasticity of 1, i.e., Constant Returns to Scale, yields similar results to 0.9, the only difference being occasional bumps on the dynamic adjustment path. Increasing Returns to Scale, as expected, lead to unstable results as the most efficient firm quickly dominates the market, and specialization depends critically on the structure of the best firms’ capabilities. For all other elasticity of production values, in both the intermediate \( b_I(i) \) and final \( b_F(i) \) production, I found that the smaller the elasticity of production, the more limited the specialization in the long-run equilibrium – which is an expected result, as non-elastic production functions curtail the potential gains from specialization. Either way, the results presented in the next section are not driven by the parametric choice of the elasticity of production.
integrates. More importantly, taking this starting point makes the results easy to follow graphically but at the same time this arbitrary starting structure (50% capacity upstream and 50% downstream for each firm) does not affect the qualitative results. I also ran the model with “unbalanced” initial endowments, starting with specialized firms, and the results again held.

Generating Scenario Implementations through Monte Carlo Simulations

Having ensured the robustness of the model in terms of parametric choices, I did a further robustness check: Using Monte Carlo simulations, I constructed 15 different implementations of each of the 3 initial scenarios – i.e., scenarios with the same parameters and capability correlations, but with different micro-structures in the productive capabilities of each firm. I tested these 15 different, randomly generated implementations to ensure that the micro-structure in firm capabilities within each scenario did not drive the qualitative results. I ensured that in all of the 15 implementations generated for each scenario, the same qualitative results held; I further confirmed that the upper and lower bounds of these 15 runs, as well as the average values, were all consistent. Within each scenario I considered 11 different TC values, and ran the model for 40 periods and 3 base conditions, and that I tried this for 15 implementations, I ran 19,800 different MCP optimization problems, each individually solved with GAMS to ensure robustness. In addition, I carried out a number of additional runs of the model, many of which are reported in the results section of the paper. In total, for this paper, about 58,000 different optimization runs were tried and recorded (including runs that were tried to simulate different entry conditions, as noted above), were consolidated by scenario, and were converted to graphs through a custom-made MS Excel database.

Model Results

Vertical Integration as a Function of Productive Capability Distribution and Transaction Costs

The first important finding is that the correlation of productive capabilities up- vs downstream is the major driver of the degree of vertical integration. A look at Figures 2a, 2b and 2c is revealing. These figures measure the degree of specialization in an industry, calculated as the fraction of the intermediate good which is traded through the market, over the total intermediate good produced.

\[19\] Specifically, I tried the model with an initial unbalanced capacity endowment (75% downstream/25% upstream for some firms, and the inverse for other firms). I considered three versions: i.e. a setup where capacity distribution was correlated with capabilities (in the sense that the most efficient firms in one segment had a higher capability in that segment); another where there was no correlation; and a third where there was, perhaps perversely, a negative correlation between capacity and capability in the initial period. For all the resulting \(3 \times 9 = 27\) graphs, representing 297 runs of 30 periods, the qualitative results discussed in the paper still held, although the graphical results were more difficult to follow. Thus, the choice of a vertically balanced initial capacity structure was made on the grounds of being both analytically plausible and consistent with the literature, and because it yielded a simpler illustration.
An index of 1 indicates full specialization (firms specialize either up or downstream), while 0 indicates full integration (no intermediate good is traded in the market).

In Figure 2a, productive capabilities along the industry’s value chain are not well correlated: some firms are weak upstream and strong downstream, or vice versa. Given such imbalances, vertical specialization does occur; furthermore, given such capability dispersion along the value chain, the reduction in TC leads to increasing levels of dis-integration. Even with substantial TC (40% of the value of the good produced), vertical specialization does emerge, despite the fact that in the initial period, upstream and downstream capacity endowments are set to be equal for all firms, and hence firms might have a reason to remain vertically integrated. We also see that for vertical specialization to start emerging, TC have to fall below a critical level, and that is the level at which the “tax” that TC imposes is offset by the benefits from accessing superior productive capabilities. In general, the threshold in terms of TC where specialization starts emerging is a function of the magnitude of the capability differences between firms; given the extreme differences in productive capabilities, bigger than those reported here, specialization occurs even with very high TC levels.\(^{20}\)

In Figure 2b, we have zero correlation between upstream and downstream capability. Comparison with Figure 2a shows that both the level and rate of increase of specialization are lower. Furthermore, the critical TC level below which specialization occurs is substantially lower than in Figure 2a. All these effects can be explained by smaller (average) capability dispersion within a firm, the attendant lack of opportunity for co-specialization between any two firms in the industry, and thereby the reduced motivation to trade.\(^{21}\)

In Figure 2c, by contrast, productive capabilities are evenly distributed along the value chain. While the overall level of capability dispersion in the industry has not changed compared to 2a, in this scenario there is a very high correlation between upstream and downstream competence – so essentially there are no trade gains to be had from specializing. So even where TC are low, no specialization occurs, because there is no strategic logic to support it.

\(^{20}\) The intuition is clear: TC and capability asymmetries along the value chain are the two opposing forces that co-determine specialization. Given sufficiently asymmetric capabilities, even extreme levels of TC will allow for specialization. However, a reduction in TC will encourage more specialization in all cases except the degenerate case of full (100%) specialization, where reduction in TC cannot lead to more specialization.

\(^{21}\) Note that in the model, specialization occurs through the choice of firms that could be integrated, yet choose not to be so. This general case can also encompass the possibility of “pure vertical specialists,” in that a vertical specialist is a firm with a good capability in one part of the production process and none (or, more strictly, very low capability) in another. So the existence of vertical specialists is consistent with the model’s setup.
By comparing Figures 2a, 2b and 2c, we can see that the main driver of vertical specialization on the aggregate level, which is the sum total of the individual decisions of specific firms, is the nature of the capability differences, and in particular the correlation of firms’ productive capabilities along the different value chain segments. It is also evident that TC reductions do play a role, but only when there is an underlying heterogeneity among the productive capabilities along the value chain; in other words, TC catalyze scope through the underlying differences in productive capabilities. Similar to the theory of international trade, where it is comparative advantage that drives the patterns of trade, here comparative advantage drives the patterns of vertical specialization. And much as international tax and transportation costs operate on underlying productivity differences to shape international specialization (Ricardo, 1817; Deardorff, 1980), TC, be they driven by asset-specificity, hold-up, coordination or friction, operate on capability differences along the value chain to determine vertical specialization.

Figures 2a and 2b also highlight the dynamic results. If we look at the evolution of specialization over time, we observe that it increases dynamically. The importance of capability differentials is magnified as efficient firms expand in their areas of strength, increasing both their share of the market and, as a result, the average specialization in the industry. This is evident from the upward slope of the specialization plane over time.

A second interesting observation is that this dynamic effect of increased specialization is more pronounced at lower TC levels; that is, low TC not only allow for a higher specialization in each period, but also enable greater rates of increase of specialization over time. The reason for this is that lower TC allow firms that are strong either downstream or upstream to expand more (at higher TC, such expansion would be less profitable, because it involves higher TC costs). Thus, the mix of production is increasingly tilted towards the specialized firms, and this specialization increases more quickly over time as the integrated firms in the industry shrink their production, and the specialized producers, who are more effective, replace them. Figure 2a in particular shows that over time, there is a significant degree of specialization. This is because as time goes by, through competition and selection, the capability pool in the industry becomes more skewed: firms that are good only up- or downstream grow proportionately more when TC are low, and as a result specialization increases as the vertically balanced firms are driven out of the market. So vertical specialization increases as a result of competition; and this increase in specialization is greater when TC are low.

Thus, the model demonstrates that gains from trade determine specialization in an industry. From a focal firm’s perspective, if its potential transactors have similar productive capabilities, then
even small TC will be sufficient to deter them from using the market. On the other hand, if firms have widely varying productive capabilities, and in particular if their capabilities are uncorrelated or are negatively correlated along the value chain, then even high TC are not sufficient to offset the economic desirability of using the market, i.e., from having a vertically specialized industry structure. This leads to the second major and perhaps the most important conclusion from the model. Specifically, a reduction in TC is the catalyst, but not the ultimate driver, of vertical disintegration. (cf. Demsetz, 1988). If we compare Figures 2a, 2b and 2c, we observe that a decrease in TC in the presence of asymmetric capability distribution (where upstream and downstream productive capabilities are negatively correlated) leads to significant specialization; however, if productive capabilities are symmetrically distributed, then a reduction in TC will not produce any effect. In other words, reduced TC lubricates the workings of the market, and allows firms to capitalize on their productive capabilities and relative strengths. If firms look different, TC reductions will allow for significant specialization, as each firm will focus on its area of strength; however, if all firms look alike, reduction in TC will do little (if anything) to promote specialization and dis-integration. There will be no latent gains from trade to change vertical scope. The insight, then, is that capability distributions (in particular, the correlation of productive capabilities between different parts of the value chain) and TC jointly determine vertical scope.

Another important observation is that vertical specialization is not driven by absolute, i.e. competitive advantage. Recall that even in the “symmetric productive capabilities” scenarios, some firms are stronger than others in a particular part of the value chain, in that they have absolute cost/capability advantage over others. Yet specialization does not occur in these cases, even though some firms have absolute advantage and others dis-advantage. The reason is that the decision to specialize hinges on the relative capability differences along the value chain, i.e. comparative advantage, rather than on absolute differences in capability levels. This suggests that the near-exclusive emphasis of the Resource-Based View (RBV) on “sustainable competitive advantage”, itself a strong form of competitive advantage (cf. Combs & Ketchen, 1999), may be misleading, at least with regard to decisions about vertical scope that are determined by simple comparative advantage.

The dynamic relationship between TC and capability distribution also makes strategic sense. As industries evolve, the rate of increasing specialization is partly a result of the pure forces of selection. That is, the increasing specialization over time in Figures 2a and 2b is not due to any exogenous change in TC; rather, it is due to the fact that specialized firms take an increasing share of the production because they are more competitive. The relative speed of increase of specialization is significantly greater for low TC values. This is because lower TC enable greater
efficiency in “natural selection,” which means that lower TC create a “virtual circle” of their own, which promotes specialization both statically, and dynamically, through competitive forces. So it may be that the gradual increase in specialization at the industry level is not related to any changes in the transactional environment, but rather to the impact of the “hand of selection”, which operates more freely in dis-integrated environments, with specialization endogenously begetting more specialization.

**Economies of Scale and Vertical Specialization**

The model allows us to confront another thorny issue – the relationship between vertical specialization and economies of scale. Interestingly, in all three of the base-case scenarios, once we include differential scale (a downstream sector whose EOS is lower than those of the upstream), specialization emerges, consistent with Stigler’s (1951) hypothesis – even though, as Figures 3a-3c show, such specialization is fairly limited. However, some more specific findings on how scale interacts with capability differences reveal an even more interesting picture.

First, by comparing Figures 3a and 2a, we can see that differential EOS lead to a smaller aggregate level of differentiation if the initial capability endowment is asymmetrically distributed. That is, if efficiencies upstream and downstream are inversely correlated, then the additional existence of relative dis-economies on only one segment reduces the aggregate degree of specialization. Why is this so? The intuition, although surprising, is convincing: if there are dis-economies in one segment (downstream), then firms with a comparative advantage downstream cannot ratchet up their advantage by growing because of these very diseconomies; which means that they will end up becoming more like the average firms in the industry; which in turn means that there will be less incentive for them (and the other co-specialized parties) to promote vertical specialization. So, in contrast to the “baseline expectation” that EOS drive specialization, under some conditions EOS may lead to more or actually less specialization, depending on the correlation between the firms’ capabilities. Note that this “specialization reducing” effect is more pronounced in the dynamics; the difference between Figures 2a and 3a is not only in terms of what happens in the initial period, but even more so in terms of how much specialization can progress, as the growth of effective firms is hampered by scale.22

---

22 It should be noted that this is the only set of scenarios where the conditions of entry matter for our results, even though even here the logic in our analysis is preserved. Specifically, if we do allow entry, then specialization can be supported by more vertically specialized, smaller units downstream, entering and remaining small (since becoming larger would mean that they would lose their relative advantage). This dynamically promotes specialization, even though even with entry, differential EOS do dampen specialization given negative or zero correlation in capabilities.
In contrast, the positive correlation scenario (Figure 3c) shows that differential EOS did produce some vertical specialization, for very low TC levels. The logic of such specialization is compatible and qualifies Stigler (1951): As production \textit{at the level of the firm} grows (and it grows more for the efficient firms), then firms (even if endowed with superior capabilities) become \textit{comparatively worse} in the segment with dis-economies of scale. Thus, firms that are equally good upstream and downstream find it profitable to grow their scalable upstream division, but not their downstream division. As a result, these firms turn to other firms for vertical co-specialization, and do find willing trading partners who cannot shield their inefficient upstream operations, but can salvage their downstream since dis-economies of scale work in their favor. Thus, specialization emerges as firms become \textit{de facto} unequal in their productivity up- vs downstream.

This finding, though, suggests that differential EOS can only lead to specialization in the presence of \textit{absolute capability differences} which lead to the potential gains from specialization, so that the Stiglerian hypothesis rests on an unstated assumption. To confirm this, I ran a scenario with \textit{absolutely identical} firms, and different upstream vs. downstream EOS. As hypothesized, \textit{no specialization} occurred – so differential EOS by itself is not a sufficient condition. This interaction between EOS and capability differences may explain why in real life industries differences in EOS are sometimes discussed as being the reasons for specialization, whereas in analytical models with firms with identical capability structures, scale is not found to be a sufficient justification for integration (cf. Dufeu, 2004). More important, our findings demonstrate that the impact of EOS on specialization depends on the underlying capability structure, and TC.

\textbf{Learning Curves and other Endogenous Capability Development Mechanisms and Scope}

The next set of results relates to the impact of learning curves. First, in Figures 4a-4c we consider how integration evolves as a function of capability distribution and TC, when both parts of the production process are subject to learning curves. Comparing Figures 4a and 2a we observe that learning curves do not have much effect on scope when the initial capability differences are strongly negatively correlated. Also, a positive correlation in capabilities (Figure 4c) will not lead to specialization, regardless of TC level, if learning curves apply both upstream and downstream. What is interesting is that learning curves \textit{do} make a difference to the way in which random capability distributions affect specialization. Figure 4b shows considerably more specialization than Figure 2b; the reason here is that productivity differences between the upstream and downstream segments are magnified by the learning curves and, over time, firms that were more or less similar in terms of their upstream vs downstream capabilities, become increasingly polarized; capability in one segment begets more production in that segment, which begets more
productivity through capability improvement, such that an initial random distribution of capabilities, through the production and the selection processes, becomes increasingly systematic, with firms being good either upstream or downstream. This process of generating increasing gains from trade is actually *fuelled* by low TC, which allow these benefits to be generated, with specialization begetting more specialization.

![Include Figures 4a to 4c and 5a to 5c about here](image)

Some even more intriguing results occur when we consider learning curves in only one of our vertical segments – e.g., only the upstream segment, shown in Figures 5a-5c. Starting from zero-correlation up-vs downstream, we see that the existence of learning curves in only one of the segments leads to even more specialization, for any level of TC, than when learning curves occur in both segments. The reason for this is simply that the existence of learning curves in only one segment creates an even greater disparity between the up- and downstream segments. Firms that are reasonably good in both segments will become even better in one of the segments, but not in both; and this will tend to enhance their advantage in specializing. The impact of learning curves occurring in only one segment is even clearer when we compare Figure 5c with either 4c or 2c. In this case, even with a symmetrical capabilities distribution, where firms that are good upstream are also good downstream, specialization emerges dynamically, and fairly rapidly. The reason for this is that the better firms produce more, and become even better on one of the two segments, so that they develop comparative advantages that focus on some of the segments; concentration, as expected in the presence of a learning curve, increases in one segment, but not in the other, and as a result, specialization increases. Thus, *dynamic unevenness along the value chain* becomes a driver of specialization. This perhaps is the dynamic described by Smith (1776) and Young (1928), and more recently by Stigler (1951); it is the existence of “dynamic EOS”, i.e. learning curves in one of the parts of the value chain, which can lead to vertical specialization by dynamically creating gains from trade. Indeed, this analytical formulation seems to echo the verbal theorizing of Langlois and Robertson (1995).

A fairly similar picture emerges if we consider other endogenous capability development mechanisms, such as those proposed in our model. Figures 6a to 6c show these endogenous rules of capability development, where profit is re-invested and can improve a firms’ capabilities; in these figures, each of the segments improves at the same pace. What we observe here is that, as in the “base case scenario”, specialization dynamically begets more specialization, but only if there is underlying capability heterogeneity; and that the effect is more pronounced at lower levels of TC; and that endogenous capability development reinforces the dynamics of specialization. This
is because over time, more successful vertical specialists become even more effective and, as such, specialization dynamically increases. Also, symmetrical endogenous capability improvement cannot produce any effects if capabilities are initially positively correlated.

As would be expected, asymmetric returns to profitability re-investment (i.e. ability to improve only one part of the value chain), lead to even greater specialization, since the more effective firms become increasingly better in one of the vertical segments. For this reason, Figures 7a to 7c exhibit remarkable degrees of vertical specialization as a result of capability improvements. Indeed, random initial distributions in capabilities are shortly driven into much more systematic ones, as the firms that are profitable de facto become the leaders in the “improvable” upstream segment. The extent of resulting specialization is noteworthy, in particular in the presence of symmetrical capability distributions (Figure 7c). The ability of firms that are better, in absolute terms, to generate profits that are then ploughed in leads to a substantial capability improvement on one sector, which leads to even greater profitability, which then invites specialization, which begets profitability, which begets more specialization and as a result lead to dramatic specialization increases almost regardless of initial conditions, even in very high TC levels.23

Extensions: Endogenous changes in Transaction Costs, Focus Benefits, Limits to Expansion

In addition to the simulations reported above, I tried a number of alternative specifications, and some extensions of the model. Below I briefly discuss some of the more relevant results.

First, endogenous changes in TC. Rather than just relying on the different values of the TC “net tax”, I tried various versions of endogenously changing TC, drawing on the literature. The first extension was to consider the potential mitigating factor of the “small numbers problem”, i.e. the additional risks of depending on a few potential suppliers of intermediate goods over and above the “base” net TC tax (Williamson, 1985). To do this, I computed a dissimilarity index of the

23 The profit-based endogenous capability development scenarios raise some fascinating issues that go beyond the confines of this paper, since they require we understand not only scope, but also profitability dynamics. Still, a brief discussion is called for. In the symmetrical capability development case (Figure 6b), for instance, the initial random differences are not exacerbated by profit-driven capability development quite as much as by the learning-curve induced ones (Figure 5b). The reason is that in the zero-correlation scenario, profitability does not accrue to the specialized firms quite as much; which then means that the balanced firms do get the opportunity to improve, in relative terms; so specialization does not beget specialization as easily as in the volume-based learning scenarios. Perhaps more intriguingly, in the asymmetric capability re-investment scenario (Figures 7a-7c), greater specialization eventually emerges in the (initially) highly correlated capability structure (Figure 7c) than in initially negatively correlated one (Figure 7a). This remarkable result happens because in the symmetrical capability scenario (Figure 7c), some firms are very profitable (best in both segments) and others not at all; this wide disparity generates substantial profitability differences. The very profitable firms can only improve one segment, in which they become increasingly strong. So over time, the scenario that was originally balanced becomes very imbalanced. Thus dynamic conditions can dramatically change or even reverse initial distributions through profit re-investment, changing gains from trade and thus specialization. These profitability-linked issues are explored in a companion paper (Author et al, 2005).
concentration upstream compared with concentration downstream; relative symmetry (e.g. same up- vs downstream distribution of power) was expected to lead to safer transacting than a dissimilar set of TC where the selling or buying side was much more concentrated. This formulation produced the expected effects, reducing specialization in the intermediate values of TC, and dampening the dynamic increase of specialization growth. However, the introduction of such endogenous TC rules did not reverse any of the findings I reported earlier.

The second form of endogenous change was to introduce systematic “learning”, and create a “learning curve” for transacting; the more that is transacted in an industry, the smaller the “net TC” becomes (Mayer & Argyres, 2004). This led to an expected increase in specialization, but the effects were only visible if there was an underlying reason to specialize; and the effects were stronger at lower or intermediate levels of initial TC. Finally, my experimentation with dyad-specific TC, or TC that could be reduced by investment of the parties, showed that the qualitative logic of our analysis was still valid. Such TC conceptualizations are left for follow-on research.

The second extension was to examine “focus benefits”; that is, benefits that would accrue to a firms’ capabilities as a function of its specialization; i.e. a firm would increase its capabilities in the segment where it focused its production, as a result of focusing its production. The results were in the expected direction; even with a focus benefit, symmetrical initial capability distributions did not yield any specialization, regardless of TC. For negatively correlated capabilities, specialization was enhanced (when compared to the “base case” of Figure 1a), and the impact of focus was greatest in the “random” case, since the existence of a “focus benefit” exacerbates the underlying capability differences statically and dynamically.

The third extension was to look at the dynamics of expansion. In particular, I considered how specialization would emerge as a result of “differential ability to expand”; for instance, of having a limit on the growth of one part of the value chain, that was more stringent than on the other. For instance, manufacturing (upstream) might be much easier to expand than distribution (downstream). Under these conditions, we can see from Figure 8 that even for positively correlated capabilities, specialization does emerge, for a purely dynamic reason. Similar results were obtained from another, related extension, i.e. the imposition of “growth costs” or “diseconomy” in expanding (cf. Rubin, 1973), due, e.g., to the fact that a firms’ success might be based on a non-replicable, non-extendable routine or recipe (Winter & Szulanski, 2001).

Contrary to the results from 2c, Figure 8 shows that vertical specialization can temporarily emerge even in the absence of capability heterogeneity along the value chain. So the unevenness of the...
expansion conditions along the value chain can, in and of itself, cause vertical specialization. This result has some interesting parallels with the results of differential scale. When the expansion costs are the same in both parts of the value chain, firms will expand (or shrink) in both segments, and remain integrated. However, if they can expand costlessly (or have the ability to grow more) in only one segment, with expansion in the other segment being costly (or limited in terms of how much growth is feasible), then expansion will be uneven: Efficient firms will expand in the segment that is easier to grow (e.g. upstream) and for the downstream they will rely on other firms; these other firms, in turn, find that they cannot afford to operate in both segments (as they are facing competition there by their more efficient rivals) but can still maintain a presence in the segment where they are shielded by competition (since the efficient firms cannot grow quickly or easily).24 This also leads to a pronounced pattern of the more efficient firms being partially integrated, i.e. relying both on in-house production and on the market, a phenomenon observed in the real world much, much more than theory would lead us to expect (Harrigan, 1985).

The role of “excessive demand”: How dis-integration slowly emerges through competition

There is one last extension that is of interest, and that is to impose a very substantial increase in the level of aggregate demand (shifting the demand parameter). The results of this are most interesting (and visible) when we consider the negatively correlated capabilities scenario. So, except for higher initial demand, Figure 9 is identical to Figure 2a. The intriguing finding is that in the scenario of Figure 9, the industry goes through a stage of vertical integration, before it eventually specializes, and only later becomes similar to Figure 2a. The question then is, why does this occur? What leads to this qualitative difference and the initial integration? The answer is that high demand means that the total capacity of all the firms in the industry can be gainfully employed. Although there is a substantial underlying capability heterogeneity, even the least efficient firms are busy making money, and given the excessive demand (which translates into insufficient capacity), the industry remains specialized, and firms expand both their segments. Yet as capacity expands, and as competition intensifies, then slowly the very weakest upstream firms (that are strong downstream) find that it may be profitable to specialize in their area of strength, at

---

24 This provides an interesting and unexpected modeling insight – namely, that vertical specialization and M&A may be substitutes. Recall that our model does not explicitly deal with M&A; if a firm wants to expand, it has to do so through organic growth, which is subject to the same limitations that organic growth often is in practice. It cannot just go and buy the resource endowment of other firms in the industry. Given that this is fact, if there are asymmetric growth limits, effective firms (even if they can grow part of their operations easily) have to rely on the capability and resource endowments of other firms that already have their operations in place. So a firm has to resort to vertical specialization, and use others’, say, downstream goods or services as it cannot grow them itself. If it had the opportunity of buying these resources needed for downstream growth, it would possibly do this to substitute for engaging in specialization. This seems to be happening in the banking sector where, either way, distribution is more difficult to grow than “production”: European regulations hampering M&A (at least when compared to the US) have led to a greater extent of vertical specialization, with large institutions collaborating with local distributors.
exactly the time that, and because, the firms which are strong upstream start to consider the benefits of finding a partner, and abandoning their inefficient capacity. As capacity increases, an increasing number of firms, with less pronounced latent gains to trade, find it profitable to co-specialize and abandon one of their vertical segments.  

The process of “rationalization” and value chain specialization that we observe in many industries (see Jacobides & Winter, 2005) may actually be due to such dynamics, whereby competition brings underlying heterogeneities to the surface and thus leads to specialization – the extent of which will then depend on the level of TC. Likewise, if we introduce a sudden drop in demand, in Figure 9 or indeed any other scenario with underlying capability heterogeneity, we observe a step-increase in vertical specialization, for the very same reason: Competitive pressures force the system to rationalize. This bears interesting similarities to North’s (1986) and Silver’s (1984) discussions of how vertical specialization emerges as a result of a crisis which leads firms to consider how to best employ their capabilities and capacity.

Discussion

This paper set out to tackle three tasks through a systemic model and its attendant numerical simulations: First, to analyze how capability distributions and scale, both independently and jointly affect integration; second, to examine how TC and capability differences interact to shape scope; third, to examine how dynamic factors such as learning curves, endogenous capability development through re-investment, and limits to expansion, catalyzed by TC, affect scope and its evolution. It has provided a thorough systemic analysis that allows us to consider how vertical scope is set in the context of competitive interaction and industry evolution. Several insights can be identified from this model, which I will briefly discuss and relate to extant theory.

First, I suggest that rather than considering TC and productive capabilities as additive factors, which have to be taken into account by an individual firm, we need to look at them systemically; specifically, we have to look at the distribution of productive capabilities along an industry’s value chain. The variability of capabilities at the industry level determines the potential gains from trade that drive the desire of firms to specialize vertically, and as such need to be studied.

Second, the analysis also shows how TC operate on this distribution of productive capabilities, shaping vertical scope. One way to visualize this is to consider a lake, and think of the level of water in the lake as being the TC level. As the level recedes, if the landscape is rugged – that is, if

---

25 The onset of specialization can, of course, be speeded up by capability development rules, or more focused or strategic growth in capacity, with firms choosing to grow the areas in which they are best.
the productive capabilities are asymmetrically distributed – islands of specialization will increasingly appear. However, if the underlying surface is smooth, if there is no heterogeneity beneath the surface to be uncovered, the landscape will not change, and integration will remain.

Third, the analysis has shown that by creating a system that “adds up”, the role of factors such as scale and capability differences produces different results from what the theory would lead us to expect. In particular, the model provides some novel insights on how scale affects integration, showing that for differential scale to have an impact there must be some underlying absolute capability differences; that the impact of differential scale depends on the correlation of capability differences, i.e. it can either attenuate or enhance vertical specialization; and that the effect of scale is moderated by TC.

Fourth, the model highlights the importance of dynamic factors, such as learning curves or capability improvement dynamics. It demonstrates that such dynamic factors tend to exacerbate random differences in the capability structure of an industry, dynamically amplifying them, thus leading to specialization. It also shows that one of the most potent drivers of specialization is the existence of asymmetric dynamic effects, whereby only one of the segments is subject to improvements or learning. This finding may offer an alternative explanation to Smith’s (1776), Young’s (1928) or Stigler’s (1951) reasoning, which seem to focus more on differences in learning curves and in the ability to develop knowledge, than on differences in scale.

These conclusions have significant implications, as they refine our view about what really affects vertical integration. Economic theory (including TCE), until the middle of the 1990s, for the most part, de-emphasized productive capability differences (Nelson, 1991). However, recently, there has been increasing interest in exploring the impact of heterogeneity on production costs, an interest that is both substantial and persistent (Lieberman & Dhawan, 2001). On example of this is the surge of interest in global outsourcing (Feenstra, 1998; Grossman & Helpman, 2002), which acknowledges the role of gains from trade and capability differences. This model has expanded on how gains from trade interact with TC, learning and capability development, and has yielded a number of new results, which show that, important as they may be, TC can only operate on underlying capability differences (Jacobides & Winter, 2005). This should contribute to a more robust understanding of scope, and complement recent efforts of to explain how and why scope evolves over time, with industries shifting from integration to dis-integration and back again (Fine, 1998; Christensen et al., 2002; Langlois, 2003; Jacobides, 2005).

In addition, the finding that vertical specialization is a function of comparative as opposed to competitive advantage, suggests that the emphasis in RBV to seek and examine only competitive
advantage and the resources and capabilities associated with it (Barney, 1986), may perhaps be misplaced. In particular, that focus may overlook some economically important drivers of vertical scope, which depend on comparative advantage and the distribution of productive capabilities, as opposed to competitive (absolute) advantages. Indeed, there is little, if any literature that unequivocally considers how productive capabilities are distributed within a segment, and how they are correlated between segments; explicit study of these issues holds much promise.

In terms of analytical structure, this model also represents a shift from the analysis of the transaction, to the analysis of the system of productive capabilities and TC as they co-evolve. The relationship between such a systemic analysis of capabilities and TC as they determine scope and division of labor, and the micro-analytic inquiry of TCE, are roughly analogous to the relationship between a general equilibrium inquiry, that acknowledges the causal interdependence of price in different markets and the study of price determination in a single market. The advantages in terms of feasibility and focus of addressing this narrower problem are enormous, but the broader problem provides a more satisfactory account of the causal logic (Jacobides & Winter, 2005). For instance, our model shows that every marginal “make-vs-buy” decision made by individual firms, depends on the relative prices that can be achieved through the intermediate market (to which TC are added) as compared to the price that the downstream division can pay; and that these two sets of prices depend, in turn, on the productive capabilities of all participating firms (not just any two firms), in both the upstream and the down-stream segment. It also shows how these capability conditions change over time. As competition changes the pool of productive firms, and shifts production to the more efficient firms, e.g. promoting specialization, the marginal “make-vs-buy” calculus conditions change as a result. These dynamics, that ultimately determine scope, require us to shift our focus to the industry, largely abandoning closed-end solutions and modeling dynamic systems instead.

Framing the problem of vertical scope as one determined by industry-wide competitive dynamics, which depend on productive capabilities and TC, also has practical implications. In particular, it explains certain theoretical puzzles, such as why reductions in TC do not always lead to less integrated firms. The reason for this may be that TC operate on capability distributions, and that capability distributions differ across settings, industries, and countries. Another implication emerges from the analysis of the dynamics, and in particular of the vertical specialization which emerges as a consequence of expansion costs. The model provides an empirically motivated explanation for a “mixed integration” strategy – that is, the concurrent use of both the firm and the market, even in equal measures, which is an anomaly for the extant theory (cf. Harrigan, 1985). It does this by combining the insights of Coase (1937) with the analysis of dynamics of capacity and
capability expansion pioneered by Penrose (1959).

Specifically, the model predicts that mixed vertical strategies will be pervasive in the presence of differential scale as well as scalability, along an industry’s value chain. Qualitative evidence from a number of sectors that I have had experience of corroborates this finding. Banks, for instance, try to sell their banking products both through their own branches, and through intermediaries. The motivation for this, I was told, is that it is much harder for banks to expand their retail branches than to expand their financial product manufacturing and customer servicing operations. Therefore, they make use of all of their internal production and, when it becomes too costly for them to expand their retail branches, they make use of “the market” in the guise of brokers, other banks, or financial intermediaries.

The model also suggests that vertical integration is affected by history and path-dependency. This is not only because of history, reputation or path-dependency in contracting (Argyres & Liebeskind, 1999), but also because of path-dependency and historical accident in productive capability distributions, which, as the model suggests, drive scope. This also explains why the vertical structures in similar industries may differ markedly across locales, markets, countries, or time periods. When the patterns of distribution of productive capabilities up- vs. downstream vary, vertical specialization also varies. This observation has another significant corollary: differences in vertical scope should not lead us to infer differences in TC; differences in productive capability distributions may well be the drivers of vertical scope. Additionally, this model underscores the role of the correlation pattern between capabilities along the value chain, and as such points to the need for and potential promise of research to understand why in some sectors there are negative, zero, or positive correlations.

The structure of the model also allows us appreciate the competitive rationale underpinning industry evolution. For instance, the explanation of why industries start off vertically integrated and only when capacity increases does specialization occur, seems to have a certain face validity, and also helps us to gauge how and why competition (through capacity expansion) interacts with TC and capability differences to dynamically determine scope. Likewise, the dynamics of increasing specialization over time, as the result of selection forces alone, show how scope may be affected not only by TC, but also by an industry’s competitive dynamics.

Such “mixed strategies,” pervasive in a number of sectors, have not been satisfactorily explained to date. While some arguments have been made in favour of both making and buying as a competitive weapon (Porter, 1980), the extensive use of both integration and market procurement cannot be readily explained by the existing research. Also, TCE does generally assume that there is “one best way” of organizing a particular transaction, and hence that firms should either make, or buy, or ally, but not engage in a mix of such solutions (Williamson, 1985, 1999). Thus, the model provides a novel rationale for tapered integration (Harrigan, 1985).
Another advantage of this model, which due to space constraints will be explored in a follow-on paper, is that it yields specific predictions about the *profitability impacts* of changing TC under different capability distributions and, as such allows us to consider how TC operate on the underlying capability distribution, statically and dynamically, to affect the level and distribution of profits in an industry, as well as the division of rents between firms and the resources they employ (see Author et al., 2005).

**Limitations, Extensions and Concluding Remarks**

This paper has several limitations. Some are inimical to any analytical model, in that it only represents a stylized, and as such not a fully realistic or complete depiction of reality. Other limitations relate to the model itself, the most important one being that the model is solved through computational, as opposed to closed-end methods. The model provides a sense of the interdependencies between different variables, and their evolution under different scenarios, but not a closed-form solution of a specific game. This is the result of both the complexity of the setting I set out to model, and of the desire to maintain some plausible behavioral foundations for the model’s analytical structure. That being said, some of the particular findings of the model could be captured and supplemented by analytical proofs, even though modeling capability heterogeneity in a system that “adds up” where most decisions are endogenous has almost inescapably got to rely on computational solutions.

While several permutations to the base structure of the model were considered and reported in the paper, clearly much more can and should be done. In future research, I plan to take the current version of the model as a reference point, and examine how changing any one dimension (e.g., foresight; pricing behavior; demand structure; etc.) affects all the relevant variables. I thus plan to introduce firm- or dyad- specific TC; increase the extent of strategizing, and model the impact of changing the capability development rules; allow for oligopolistic behavior when consolidation occurs; introduce endogenous margin-setting; allow for strategizing in the markets for resources; model for heterogeneity in resources; consider multiple types of intermediate goods, with and without TC, all of which are feasible extensions of this flexible modeling structure.

Another limitation of the model is that it considers TC and productive capabilities as distinct categories. While this yields some tractable results, future modeling should allow for TC choices to directly affect productive capabilities. We should also consider how capabilities are themselves shaped by TC, and examine the resulting vertical scope over time (cf. Jacobides & Winter, 2005). In particular, an extension of the model would be to allow firms to invest in reducing TC, thus
changing their institutional environment. Finally, this model could also be empirically tested, e.g. by measuring capability differences and considering how they interact with TC.

These limitations aside, the potential contribution of this paper is three-fold. First, it represents the first formal analysis of how capability differences drive integration. Taking firm heterogeneity seriously, the model provides explicit predictions about how productive capability differentials affect vertical scope, pointing to relative rather than absolute advantage. The results suggest that the distribution of capabilities is a major driver of scope, and as such require to be studied: When is it that capabilities are symmetrically distributed along the value chain, and when / why is it not?

Second, and more importantly, the model combines and integrates the analysis of (a) capability differences; (b) EOS; (c) TC, as well as (d) purely dynamic elements such as learning curves or capability improvement dynamics (and even endogenous TC changes), whose impacts have not been considered to date. A key contribution of this paper is in bringing these four elements together, to form a consistent body of theory. The resulting integrative theory and analytical approach can also address, using the same “toolkit,” both the question of determining vertical scope and the strategic ramifications of changing scope.

Third, the model shifts the mode of analysis from a ceteris paribus analysis, looking at one transaction at a time, to a systemic analysis of how scope is codetermined by capabilities, TC, and industry dynamics that, crucially, “adds up”. This shift helps us appreciate that we cannot really separate the question of division of labor between firms, and division of labor across the vertical divide – i.e. specialization. Also, to understand vertical scope and its evolution we need to focus on issues that have not thus far attracted attention, such as the distribution of capabilities across an industry’s value chain, and comparative advantage, the nature of learning curves and of the endogenous capability development process, or the limits to growth in different vertical segments of the industry. Indeed, the contribution of this paper lies not in its technique, but rather its implications for framing and theory.

Last, the model, which was partly inspired by recent qualitative fieldwork, could also provide a template for a history-friendly simulation (Malerba et al., 1999), and explain patterns of industry evolution. It could be used to explain the evolution of vertical scope, as well as to map the implications of changes in TC given particular capability distributions. Such “if-then” exercises on the drivers and the strategic ramifications of changing scope could benefit managers and policy makers alike, and also lead to further research and policy projects.
Appendix: The Model

Static Model Part I: Maximizing Production Profits

Each firm is composed of an intermediate good division, and a final good division, which contribute to the aggregate profit that it aims to maximize. This leads to the following optimization problem faced by every firm $i$:

$$
\max_{(R_I(i), Q_{IS}(i), Q_{IS}(i), R_F(i), Q_{IB}(i))} Q_F(i) \cdot P_F + Q_{IS}(i) \cdot P_{IS} - Q_{IB}(i) \cdot P_{IB} - R_I(i) \cdot P_{IR} - R_F(i) \cdot P_{RF}
$$

subject to:

$$
Q_{IP}(i) = a_I(i) \cdot R_I(i)^{b_I(i)}
$$

$$
Q_F(i) = a_F(i) \cdot R_F(i)^{b_F(i)}
$$

$$
Q_{IS}(i) = \phi \cdot (Q_{IP}(i) + Q_{IB}(i) - Q_{IS}(i))
$$

$$
Q_{IS}(i) \leq Q_{IP}(i) + Q_{IB}(i)
$$

$$
R_I(i) \leq R_{I\text{max}}(i)
$$

$$
R_F(i) \leq R_{F\text{max}}(i)
$$

$R_I(i), R_F(i), Q_{IS}(i), Q_{IB}(i) \geq 0$, where:

- $Q_{IS}(i)$: the quantity of the intermediate product that firm $i$ wishes to supply to the intermediate market at any given price.
- $Q_{IB}(i)$: the quantity that a firm demands from the intermediate goods markets at any given price. This quantity (which initially is not identical to that available to the firm from those selling intermediate goods) will clear through general equilibrium equation (II), which clears the intermediate goods market with an industry-level price for intermediate goods.
- $a_I(i)$: the capability of the intermediate production function of firm $i$ (updated each period).
- $a_F(i)$: the capability of the final production function of firm $i$ (updated each period).
- $b_I(i)$: the exponent for the intermediate production function of firm $i$ (constant over time).
- $b_F(i)$: the exponent for the final production function of firm $i$ (constant over time).
- $\phi$: the Leontieff production coefficient.
- $R_I(i)$: the quantity of intermediate resources used in production process.
- $R_F(i)$: the quantity of final resources used in production process.
- $Q_{IP}(i)$: the quantity of intermediate product the firm produces.
- $Q_F(i)$: the quantity of final product that firm $i$ wishes to supply to the final market for any given price.
- $P_F$: the price of the final product.
- $P_{IR}$: the price at which the firm buys the intermediate resource.
- $P_{RF}$: the price at which the firm buys the final resource.
- $P_{IB}$: the price at which the firm buys the intermediate good.
- $P_{IS}$: the price at which the firm sells the intermediate good.
- $R_I(i)$: the quantity of the intermediate resource used.
- $R_F(i)$: the quantity of the final resource used.
- $R_{I\text{max}}(i)$: the available intermediate resource endowment (also called maximum capacity).
- $R_{F\text{max}}(i)$: the available final resource endowment (also called maximum capacity).

The firm-specific production function describes the capability of turning $R_I(i)$ units of intermediate resource into $Q_{IP}(i)$ units of intermediate output, through the relationship

$$
Q_{IP}(i) = a_I(i) \cdot R_I(i)^{b_I(i)}, a_I(i) > 0, b_I(i) > 0.
$$

where $b_I(i)$ is the elasticity of production for the intermediate good. Moreover, total production cannot exceed the limits posed by the available resource endowment.

---

1 For both final and intermediate goods, we can generalize the formulation to the multidimensional case of vectors of inputs which are transformed into output; the resource price can be thought of as the price of the resource vector.
The firm-specific capability \( a_I(i) \) and the maximum capacity \( R_{I\text{max}}(i) \) are assumed to be fixed within a given period, but are updated in the dynamic part of the model. We also must ensure that the quantity of intermediate sold, is either produced or bought.\(^2\)

\[(1c) \quad Q_{IS}(i) \leq Q_{IB}(i) + Q_{IP}(i)\]

For final goods, the firm-specific production function describes the capability of turning \( RF(i) \) units of final resource into \( Q_F(i) \) units of output, through the relationship:

\[(2a) \quad Q_F(i) = a_F(i) \cdot RF(i)^{b_F(i)}, \quad a_F(i) > 0, \quad b_F(i) > 0.\]

The above structure can account for both economies- and dis-economies of scale, the former corresponding to exponents \( b_I(i) \) and \( b_F(i) \) greater than 1, the latter corresponding to exponents less than 1. The flexibility of choosing separate \( b_I(i) \) and \( b_F(i) \) allows us to study the impact of differential exponents upstream vs downstream on vertical scope.

The constraints associated with the final goods producer are as follows:

\[(2b) \quad RF(i) \leq RF_{\text{max}}(i)\]

As in the intermediate good production, \( a_F(i) \) and \( RF_{\text{max}}(i) \) are given and fixed within the period, and are updated in the dynamic part of the model. For the final good production, we also have to ensure that we have the necessary intermediate good, whether bought or produced:

\[(2c) \quad Q_F(i) = \phi \cdot (Q_{IP}(i) + Q_{IB}(i) - Q_{IS}(i))\]

with \( \phi \) being the Leontieff coefficient, which can be normalized to 1 without loss of generality.

Therefore, the optimization problem of the firm, described by equation (1), consists of determining \( R_I(i), Q_{IB}(i), Q_{IS}(i), RF(i) \) (which also determines \( Q_{IP}(i) \) and \( Q_F(i) \) through equations (1a) and (2a) ). Solving this profit-maximization problem leads firms to produce the optimal amount of intermediate goods, and to choose optimally either to sell them on the market or give them to the downstream division – or both; and similarly to choose the optimal amount of final good to be produced (if any) for any set of prevailing prices. For their individual profit maximization, then, firms take prices (for resources \( PRI \) and \( PRF \); for intermediate goods bought or sold, \( P_{IS} \) and \( P_{IB} \); and for the final good, \( PF \)) as given, and decide their profit-maximizing quantities. However, these prices are endogenous to the entire industry, as I explain below.

**Static Model Part II: Global Equilibrium Conditions Linking the Individual Problems**

I have described the separate decision-making processes of the \( i \) firms, reflected in \( i \) separate optimization problems. That is, given the price vector, each firm decides the quantities that it is willing to buy, trade and produce, in terms of both final and intermediate goods. But prices are themselves endogenous to the model, and they link together the individual firm optimization problems, as Figure 1c shows. In particular, for the model to equilibrate and a solution to emerge, four general equilibrium conditions need to be satisfied:

First, the market for intermediate resources has to clear through the price of intermediate resources \( PRI \). Total intermediate resource demand is obtained by adding up individual demands that are the

---

\(^2\) Note that the upstream part of the firm decides what it will produce and what it will purchase from the outside. In theory, this division could be to act as a broker, buying from other firms their intermediate production, and selling it to yet other firms; in equilibrium, of course, such arbitrage will not happen.
results of the firm-specific optimization problems. Supply $S(PRI)$ is price sensitive with an elasticity of supply $\varepsilon SI \geq 0$, and $SI$ is a supply-constant.

$$S(PRI) = SI \cdot PRI^{SI} \geq \sum_{i} RI(i) \to PRI$$

Second, the market for final resources has to clear through $PRF$. Total demand is obtained by adding up the individual demands that result from the firm-specific optimization problems. Supply $S(PRF)$ is price sensitive with an elasticity of supply $\varepsilon SF \geq 0$, and $SF$ is a supply-constant:

$$S(PRF) = SF \cdot PRF^{SF} \geq \sum_{i} RF(i) \to PRF$$

Third, the market for intermediate products has to clear. This market is a trade pool, i.e., buyers do not differentiate by origin, and sellers set a uniform price by destination. Transaction costs $TC$ are added to the equilibrium price as a per valorem tax and paid by the buyer:

$$\sum_{i} QIS(i) = \sum_{i} QIB(i) \to PIS$$

$$PIB = PIS \cdot (1 + TC)$$

Net transaction costs are thus seen as a net outflow from the system. This conceptualization is discussed in greater detail in the body of the paper. Although different firms can set different prices for the intermediate good, in equilibrium these have to be equal for the market to clear; that is, the implied structure of the intermediate market is competitive.

The last market that needs to clear is the market for the final product. The supply of the final good is given by summing up individual supplies that result from the individually optimal decisions of firms in the industry, described in equation (1). Demand $D(PF)$ is price sensitive, with a demand constant $DF$ and elasticity of demand $\varepsilon F$, and is shown in equation (IV) below:

$$D(PF) = DF \cdot PF^{DF} \geq \sum_{i} QF(i) \to PF$$

Parameter $\varepsilon F \geq 0$ is the elasticity of demand for the final good, $PF$ is the price of the final good, and parameter $DF$ is a demand constant. This ensures, that, in equilibrium, demand is equal to supply

$3$ The arrow indicates the price, which corresponds to the market clearing in this general equilibrium equation.

$4$ As a first extension, I endogenize $TC$ by making them a function of the concentration of suppliers in the market for the intermediate good. To this end, a dissimilarity index is set as

$$DissInd = 0.5 \cdot \frac{\sum_{i} |MarketShare(i) - 1|}{\text{number of firms}}$$

$$MarketShare(i) = \frac{QIS(i)}{\sum_{j} QIS(j)}$$

Based on the above index of dissimilarity, $TC$ are set as

$$TC = baseTC \cdot (1 + DissInd),$$

where $baseTC$ represent the “per valorem TC tax” which is the parameter varied in our scenarios.

A second extension consists of building “learning curves” for $TC$, so that they decline over time as a function of the amount of intermediate trade that has occurred. In particular, we can stipulate that

$$TC = baseTC \cdot \frac{cTC}{\log(cumQIS)},$$

with $cTC$ a scale factor and $cumQIS$ the cumulative quantity of intermediate sold.

These two endogenous $TC$ evolution effects can be had separately or jointly, through an additive function.
(prices adjust, through εF, until supply meets demand).

These four global conditions link the individual optimizations together, creating the Mixed Complementarity Problem (MCP) structure – a set of non-linear optimization problems for which solution algorithms exist in the computational general equilibrium literature.\(^5\) Therein, firms decide how much they produce up- and down-stream, as well as whether they buy or sell intermediate goods, and at what price, under given constraints, as Figures 1b-1c illustrate. In addition, the global equilibrium conditions ensure that the intermediate and final goods and resources markets clear. (See Ferris & Kanzow 2002 for a mathematical exposition; Capros et al, 1998, for applications).

The MCP structure allows us to solve these optimizations concurrently, linking each optimization problem with a set of global equilibrium variables. Analytically, the MCP problems in the model were solved using the Karush-Kuhn-Tucker first order conditions. Numerically, the static model was solved using GAMS (Generalized Algebraic Modeling System) for the KKT conditions. Specifically, I used PATH, a GAMS algorithm for solving MCPs (see Ferris & Munson, 2000).

**Model Dynamics: Updating Resource Endowments and Productive capabilities\(^6\)**

After a particular period \(t-1\) is over, firms change the factors that were fixed in that previous period. More precisely, their desired capacity endowments for the next period \(t\) in the intermediate and final goods segments are respectively

\[
\begin{align*}
(A1a) & \quad (1 + \text{CapResI}(i)) \cdot \text{RI}(i,t-1) \\
(A1b) & \quad (1 + \text{CapResF}(i)) \cdot \text{RF}(i,t-1).
\end{align*}
\]

With this “capacity-buffer” \(\text{CapResI}(i)\) or \(\text{CapResF}(i)\), the firm is prepared to e.g., meet higher demand in the next period or take over market shares from its competitors.

I incorporate learning by giving the option of adjusting the firms’ capabilities, both upstream and downstream, dynamically. The intuition behind our update rule is that a firm can learn from experience, i.e. we assume that its efficiency (represented by its capability-coefficients) increases commensurate with its cumulative output. Formally, we set

\[
\begin{align*}
al(i,t) & = al(i,1) \cdot (1 + cI(i) \cdot \log(\text{cumOutputI}(i,t-1))) \\
af(i,t) & = af(i,1) \cdot (1 + cF(i) \cdot \log(\text{cumOutputF}(i,t-1)))
\end{align*}
\]

with \(c(i)\) a scale factor that represents the “speed of institutional learning”, i.e. describes the rate at which the firm can transform experience gained from production in capability improvement. \(\text{cumOutputI}(i,t-1)\) denotes the cumulative output upstream of firm \(i\) up to time \(t-1\). Due to the logarithmic formulation, doubling the cumulative output always leads to a constant increase in capability. Accordingly, we set

\[
\begin{align*}
al(i,t) & = al(i,1) \cdot (1 + cI(i) \cdot \log(\text{cumOutputI}(i,t-1))) \\
af(i,t) & = af(i,1) \cdot (1 + cF(i) \cdot \log(\text{cumOutputF}(i,t-1)))
\end{align*}
\]

for the downstream segment.\(^7\)

---

\(^5\) In order to confirm our results, I also solve another problem, optimizing the simple NLP for the entire industry (where we maximize total industry profits). A basic lemma of general equilibrium theory (e.g., Koopmans, 1951) is that, at given prices, and in the absence of externalities, aggregate profits are maximized if, and only if, individual firm profits are maximized. This comparison of the multi-firm results against the industry-wide optimization provided further confirmation that the analytical formulation did not contain any errors.

\(^6\) In this section, we extend the notation introduced in the description of the static part to incorporate a time-index \(t\), enabling us to account for the dynamic structure of the model.

\(^7\) I also introduced, as an extension, a concept termed **focus benefit**, which can dynamically influence the firms’ capabilities. Specifically, this benefit rests on the presupposition that a firm that focuses on one segment – either upstream or downstream – can increase its capability in the segment it specializes in. Formally, focus benefit is set as
I further provide the option that the profit a firm earns in a given period has an impact on the evolution of its capability structure. Formally, if $P(i,t-1)$ denotes the profit of firm $i$ in period $t-1$, we posit that it invests

$$IP(i,t-1) = IPCoeff(i) \cdot P(i,t-1)$$

with $IPCoeff(i)$ being the investable profit coefficient (i.e. the share of profit going to capability re-investment) and $IP(i,t-1)$ the total investable profit. The next step is to decide how to allocate/pro-rate the investable profit to the two divisions – upstream and downstream. The simplest solution is to pro-rate the investable profit it to the upstream and downstream division on the basis of $QF(i,t-1)$ (quantity of final good produced by firm $i$ in period $t-1$) and $QIP(i,t-1)$ (quantity of intermediate good produced by firm $i$). Analytically, this means that firm $i$ invests

$$QIP(i,t-1) = IP(i,t-1) \cdot \frac{QIP(i,t-1)}{QIP(i,t-1) + QF(i,t-1)}$$

and

$$QF(i,t-1) = IP(i,t-1) \cdot \frac{QF(i,t-1)}{QIP(i,t-1) + QF(i,t-1)}$$

in period $t-1$ to improve its up- and downstream capabilities, respectively. Having determined the invested profits, we can now turn to the question of how investment affects capabilities. I propose a formulation that contains a fixed term (the intuition being that the absolute level of re-investment offers a new solution / idea / technique that improves productive capability regardless of scope of operation); and a quantity-adjusted component, which accounts for the fact that some of these improvements must be applied to the resources on the ground, so that the benefit cannot be expected to be related only to the absolute level of R&D, but relates also to the amount of “training” productive capacity on the ground. This would suggest that for the upstream segment, the capability of firm $i$ in period $t$ will be updated according to the rules

$$RRI(i) = aI(i,t) \cdot (1 + CIPI(i,t-1) \cdot ARI(i) + CIPI(i,t-1))$$

and equivalently, for the downstream segment, it would be that

$$RRF(i) = aF(i,t) \cdot (1 + CIPF(i,t-1) \cdot ARF(i) + CIPF(i,t-1))$$

with $ARI(i)$ being the absolute return on investment and $RRI(i)$ the relative return on investment, both for the upstream segment. $CIPI(i,t-1)$ denotes the cumulative invested profit up to period $t-1$. If both learning-based and investment-based capability update-rules are to be applied simultaneously, they have to be combined in a multiplicative fashion. This leads to

$$aI(i,t) = aI(i,1) \cdot (1 + CIPI(i,t-1) \cdot ARI(i) + CIPI(i,t-1)) \cdot \frac{RRI(i)}{RImax(i,t-1)}$$

and equivalently, for the downstream segment, it would be that

$$aF(i,t) = aF(i,1) \cdot (1 + CIPF(i,t-1) \cdot ARF(i) + CIPF(i,t-1)) \cdot \frac{RRF(i)}{RFmax(i,t-1)}$$

with $RRI(i)$ the relative return on investment, both for the upstream segment. $CIPI(i,t-1)$ denotes the cumulative invested profit up to period $t-1$. If both learning-based and investment-based capability update-rules are to be applied simultaneously, they have to be combined in a multiplicative fashion. This leads to

$$aI(i,t) = aI(i,1) \cdot (1 + CIPI(i,t-1) \cdot ARI(i) + CIPI(i,t-1)) \cdot \frac{RRI(i)}{RImax(i,t-1)} \cdot (1 + c(i) \cdot \log(\text{cumOutput}(i,t-1)))$$

$$aF(i,t) = aF(i,1) \cdot (1 - c(i) \cdot \log(\text{cumOutput}(i,t-1)))$$

with

$$c(i) = \text{MarketShare}(i,t-1) - \text{MarketShare}(i,t-1)$$

This update rule is applied to the capabilities after adjusting for scale effects (see above). Here, cFocI and cFocF are scalars that determine the impact of specialization on capabilities. MarketShareI and MarketShareF are self-explanatory.
\[ aF(i,t) = aF(i,1) \cdot (1+ CIPF(i, t-1) \cdot ARF(i) + CIPF(i) \cdot \frac{RRF(i)}{RF_{\text{max}}(i,t-1)}) \cdot (1+cF(i) \cdot \log(\text{cumOutputF}(i,t-1))) \]

In addition to these capability update/reinvestment rules, several others were considered; see footnote 11 in text.

Finally, a last type of “adjustment limit” is introduced, in that no firm can increase or shrink its capacities by more than a factor \( \Delta I_{\text{max}} \) upstream and \( \Delta F_{\text{max}} \) downstream. Each firm, then, decides its expansion policy, and the following two update rules apply:

\[ (A2c) \quad Rl_{\text{max}}(i,t) := \max\left\{ \min\{ (1+\text{CapResI}(i)) \cdot Rl(i,t-1), (1 + \Delta I_{\text{max}}(i)) \cdot Rl_{\text{max}}(i,t-1) \}, \right. \\
\left. (1 - \Delta I_{\text{max}}(i)) \cdot Rl_{\text{max}}(i,t-1) \} \right. \]

\[ (A2d) \quad RF_{\text{max}}(i,t) := \max\left\{ \min \{ (1+\text{CapResF}(i)) \cdot RF(i,t-1), (1 + \Delta F_{\text{max}}(i)) \cdot RF_{\text{max}}(i,t-1) \}, \right. \\
\left. (1 - \Delta F_{\text{max}}(i)) \cdot RF_{\text{max}}(i,t-1) \} \right. \]

where
- \( \Delta I_{\text{max}}(i) \) is the maximum relative change per period for intermediate capacity
- \( \Delta F_{\text{max}}(i) \) is the maximum relative change per period for final capacity
- \( \text{CapResI}(i) \) is the relative capacity-reserve over current production \( Rl(i) \) the firm aims at for its intermediate division
- \( \text{CapResF}(i) \) is the relative capacity-reserve over current production \( RF(i) \) the firm aims at for its final division.

**Linking Statics and Dynamics**

The final element of the model links the results of the dynamic model (specific to each firm) to the new period: maximum capacities are updated according to rule (A2); and given the new resulting capacities, the productive capabilities can be updated using (A1). Furthermore, on the industry level, demand and resource supply are updated according to

\[ (A3a) \quad S_I(t) = S_I(t-1) \cdot (1 + \Delta S_I) \]
\[ (A3b) \quad S_F(t) = S_F(t-1) \cdot (1 + \Delta S_F) \]
\[ (A3c) \quad D_F(t) = D_F(t-1) \cdot (1 + \Delta D_F), \]

with \( \Delta S_I \) being the relative change in \( S_I \) per period, \( \Delta S_F \) being the relative change in \( S_F \) per period, and \( \Delta D_F \) being the relative change in \( D_F \) per period. On the basis of these updates, the static model is run for the new period \( t \), and the cycle starts anew; statics are linked to dynamics, and we can run an arbitrary number of periods.


New York, NY: 514-530


Table 1:  
Research Design and Resulting Scenario Structure:  
Generating the Graphs

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Upstream / Downstream Capability Correlation</th>
<th>Other “treatment” Effects</th>
<th>Transaction Costs</th>
<th>Figures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>Negative Correlation (r = -0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>2a</td>
</tr>
<tr>
<td>1b</td>
<td>Uncorrelated (r = 0)</td>
<td>None</td>
<td>0-100%</td>
<td>2b</td>
</tr>
<tr>
<td>1c</td>
<td>Positive Correlation (r = 0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>2c</td>
</tr>
<tr>
<td>2a</td>
<td>Negative Correlation (r = -0.99)</td>
<td>EOS Upstream &gt;</td>
<td>0-100%</td>
<td>3a</td>
</tr>
<tr>
<td>2b</td>
<td>Uncorrelated (r = 0)</td>
<td>EOS Downstream</td>
<td>0-100%</td>
<td>3b</td>
</tr>
<tr>
<td>2c</td>
<td>Positive Correlation (r = 0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>3c</td>
</tr>
<tr>
<td>3a</td>
<td>Negative Correlation (r = -0.99)</td>
<td>Learning Curves</td>
<td>0-100%</td>
<td>4a</td>
</tr>
<tr>
<td>3b</td>
<td>Uncorrelated (r = 0)</td>
<td>Up and Down-stream</td>
<td>0-100%</td>
<td>4b</td>
</tr>
<tr>
<td>3c</td>
<td>Positive Correlation (r = 0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>4c</td>
</tr>
<tr>
<td>4a</td>
<td>Negative Correlation (r = -0.99)</td>
<td>Learning Curves</td>
<td>0-100%</td>
<td>5a</td>
</tr>
<tr>
<td>4b</td>
<td>Uncorrelated (r = 0)</td>
<td>Only Up-stream</td>
<td>0-100%</td>
<td>5b</td>
</tr>
<tr>
<td>4c</td>
<td>Positive Correlation (r = 0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>5c</td>
</tr>
<tr>
<td>5a</td>
<td>Negative Correlation (r = -0.99)</td>
<td>Capability improves (profit re-investment)</td>
<td>0-100%</td>
<td>6a</td>
</tr>
<tr>
<td>5b</td>
<td>Uncorrelated (r = 0)</td>
<td>Up and Down-stream</td>
<td>0-100%</td>
<td>6b</td>
</tr>
<tr>
<td>5c</td>
<td>Positive Correlation (r = 0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>6c</td>
</tr>
<tr>
<td>6a</td>
<td>Negative Correlation (r = -0.99)</td>
<td>Capability improves (profit re-investment)</td>
<td>0-100%</td>
<td>7a</td>
</tr>
<tr>
<td>6b</td>
<td>Uncorrelated (r = 0)</td>
<td>Only Up-stream</td>
<td>0-100%</td>
<td>7b</td>
</tr>
<tr>
<td>6c</td>
<td>Positive Correlation (r = 0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>7c</td>
</tr>
<tr>
<td>7 (c)</td>
<td>Positive Correlation (r = 0.99) (identical to scenario 1(c))</td>
<td>Limits to expansion high in downstream</td>
<td>0-100%</td>
<td>8</td>
</tr>
<tr>
<td>8 (a)</td>
<td>Negative Correlation (r = -0.99) (identical to scenario 1(a))</td>
<td>Initial Demand very high</td>
<td>0-100%</td>
<td>9</td>
</tr>
</tbody>
</table>

The dependent variable in the figures is aggregate vertical specialization, i.e. the percentage of intermediate product procured through the market.

Each graph represents one scenario of 40 periods for 11 values of TC, i.e. each figure represents the summary result of 440 optimization problems.
Table 2: Other Factors Examined – No Graphs Shown in the Paper

<table>
<thead>
<tr>
<th>Focus Benefits</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Firms capabilities improve as a function of their specialization in one segment</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Dissimilarity in the concentration of potential buyers and sellers of intermediate good leads to higher net transaction cost</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Firms learn to contract and net TC are reduced as a function of the cumulative volume transacted</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Modeled entry of firms with similar capability structures as in the base scenario; deferent demand adjustment scenarios tried. No material effect, except in different EOS</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Parametric Choices and Robustness Checks

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Range considered</th>
<th>Effects on results</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elasticity of Demand</td>
<td>1</td>
<td>0.5 to 1.5</td>
<td>Not reversed</td>
<td></td>
</tr>
<tr>
<td>Elasticity of Upstream Resource Supply</td>
<td>1</td>
<td>0.5 to 1.5</td>
<td>Not reversed</td>
<td></td>
</tr>
<tr>
<td>Elasticity of Downstream Resource Supply</td>
<td>1</td>
<td>0.5 to 1.5</td>
<td>Not reversed</td>
<td></td>
</tr>
<tr>
<td>Elasticity of Production (Returns to Scale)</td>
<td>0.9</td>
<td>0.5 to 1.2</td>
<td>Not reversed</td>
<td>Increasing Ret to Scale yield unstable solutions; CRS leads to some adjustment swings; lower values partly taper specialization cf. footnote 13</td>
</tr>
<tr>
<td>Focus benefit</td>
<td>0.02</td>
<td>0 to 0.02</td>
<td>Not reversed</td>
<td></td>
</tr>
<tr>
<td>Transaction Costs</td>
<td>0 - 100%</td>
<td>0 to 200%</td>
<td>Not reversed</td>
<td>Any non-negative TC consistent</td>
</tr>
<tr>
<td>Base Transaction Costs (used for endogenous TC)</td>
<td>0 - 100%</td>
<td>0 to 200%</td>
<td>Not reversed</td>
<td>Any non-negative TC consistent</td>
</tr>
<tr>
<td>Investable profit coefficient</td>
<td>50%</td>
<td>0 to 100%</td>
<td>Not reversed</td>
<td></td>
</tr>
<tr>
<td>Speed of institutional learning</td>
<td>0.15</td>
<td>0 to 0.3</td>
<td>Not reversed</td>
<td></td>
</tr>
<tr>
<td>Number of firms</td>
<td>21</td>
<td>5 to 40</td>
<td>Not reversed</td>
<td>Almost no change at all</td>
</tr>
<tr>
<td>Absolute return on investment</td>
<td>0.005</td>
<td>0 to 0.01</td>
<td>Not reversed</td>
<td></td>
</tr>
<tr>
<td>Relative return on investment</td>
<td>0.05</td>
<td>0 to 0.1</td>
<td>Not reversed</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1a:
(Static) Model Overview: The Basic Structure

Upstream Resource

Intermediate good

Upstream Production

Intermediate Market

Downstream Production

Final Market

Figure 1b:
(Static) Model Overview: What Firms Do Each Period

Firms: Maximize Joint Profits

Integration is an incidental result of profit-maximization

Fixed Firm-Level Variables, updated dynamically:
* Resource Endowment (Capacity constraint)
* Capability (productivity)

Fixed Industry-Level Variables:
* Per value Transaction Costs (constant through time)

Endogenous Variables – Industry Level:
* Price for Intermediate Resource
* Price for Intermediate Good (if traded)

Choice Variables
* Intermediate Quantity Produced (scale)
* Intermediate Quantity Bought from other firms
* Intermediate Quantity Sold to other firms
* Intermediate Quantity Transferred Downstream

Firm-Level Variables, updated dynamically:
* Resource Endowment (Capacity Constraint)
* Capability (productivity)

Fixed Industry-Level Variables:
* Production function (need upstream good)

Endogenous Variables – Industry Level:
* Price for Downstream Resource
* Price for Final Good

Choice Variables
* Final Good Produced (scale)
**Figure 1b:**
The Short-term (Static) Model: A Mixed Complementarity Solution

**Firm i.**

**Upstream Division**

*Maximize Profit*, by
- Selling / buying int. good
- Transferring (to division) subject to
- Capacity Constraint (resource endowment)
- Production Function given

**Downstream Division**

*Maximize Profit*, by
- Selling to final market subject to
- Capacity Constraint (resource endowment)
- Production Function
- Needed upstream input

Given:
- Downstream Resource Price

Intermediate Good Transferred

Intermediate market trade pool

\[ Q_{IS}(i) \]  \hspace{5em} \[ Q_{IB}(i) \]

\[ Q_{IS}(j) \]  \hspace{5em} \[ Q_{IB}(j) \]

Clears when sum of \[ Q_{IS}(i) \] equals sum of \[ Q_{IB}(i) \] for all firms \( i, j, \ldots \)

\[ QF(i) \]  \hspace{5em} \[ QF(j) \]

\[ RF(i) \]  \hspace{5em} \[ RI(i) \]

\[ QF = \sum_i QF(i) \]

\[ RF \cdot PRI \]

\[ RI \cdot PRI \]

Demand for final good:
\[ DF \cdot PF^{RF} \]

Supply for final resource:
\[ SF \cdot PRF^{SF} \]

Supply for intermediate resource:
\[ SI \cdot PRI^{SI} \]

\[ PRI \]

\[ PRF \]
Figure 2a: Aggregate Specialization & Transaction Costs over Time

*Negative Up / Downstream Capability Correlation*

Figure 2b: Aggregate Specialization & Transaction Costs over Time

*Zero Up / Downstream Capability Correlation*

Figure 2c: Aggregate Specialization & Transaction Costs over Time

*Positive Up / Downstream Capability Correlation*
Figure 3a: **Negative Up / Downstream Capability Correlation, Differential EOS**

Figure 3b: **Zero Up / Downstream Capability Correlation, Differential EOS**

Figure 3c: **Positive Up / Downstream Capability Correlation, Differential EOS**
Figure 4a: *Negative Up / Downstream Capability Correlation, Learning Curves in both Segments*

![Graph showing negative correlation between specialization and base TCs over periods]

Figure 4b: *Zero Up / Downstream Capability Correlation, Learning Curves in both Segments*

![Graph showing zero correlation between specialization and base TCs over periods]

Figure 4c: *Positive Up / Downstream Capability Correlation, Learning Curves in both Segments*

![Graph showing positive correlation between specialization and base TCs over periods]
Figure 5a: **Negative Up / Downstream Capability Correlation, Learning Curves only Upstream**

![3D graph showing negative correlation between specialization and base TCs over time periods.]

Figure 5b: **Zero Up / Downstream Capability Correlation, Learning Curves only Upstream**

![3D graph showing zero correlation between specialization and base TCs over time periods.]

Figure 5c: **Positive Up / Downstream Capability Correlation, Learning Curves only Upstream**

![3D graph showing positive correlation between specialization and base TCs over time periods.]

53
Figure 6a: *Negative Up / Downstream Capability Correlation, Profit re-invest in both Segments*

Figure 6b: *Zero Up / Downstream Capability Correlation, Profit re-invest in both Segments*

Figure 6c: *Positive Up / Downstream Capability Correlation, Profit re-invest in both Segments*
Figure 7a: *Negative Up / Downstream Capability Correlation, Profit re-invest only Upstream*
Figure 8: *Positive Up / Downstream Capability Correlation, Limits to Expansion High Downstream, Low Downstream*

![Figure 8: Positive Up / Downstream Capability Correlation, Limits to Expansion High Downstream, Low Downstream](image)

Figure 9: *Negative Up / Downstream Capability Correlation, Initial / Aggregate Demand Very High*

![Figure 9: Negative Up / Downstream Capability Correlation, Initial / Aggregate Demand Very High](image)