HOW CAPABILITY DIFFERENCES, TRANSACTION COSTS 
AND SCALABILITY INTERACT TO SHAPE VERTICAL SCOPE

MICHAEL G. JACOBIDES 
Assistant Professor of Strategic and International Management
London Business School
Sussex Place, Regent’s Park, London NW1 4SA, United Kingdom
Tel (011) 44 20 7706 6725; Fax (011) 44 20 7724 7875
mjacobides@london.edu

I would like to thank Sid Winter for his guidance and for shaping the ideas of this paper; Takis Georgakopoulos for his advice in modeling and GAMS; Stefan Kassberger for outstanding programming assistance and advice; and Bruce Weber, Phanish Puranam and David Croson for useful comments. This research was partly carried out while the author was a NATO Science Fellow. Funding by the Centre for the Network Economy, the Leverhulme Trust Digital Transformations Research Programme and the Research and Material Development Fund at the London Business School is gratefully acknowledged. Participants of the Reg. Jones Strategy Brownbag at the Wharton School, as well as seminar participants in LBS, INSEAD, CMU, UT Austin, CESPRI / Bocconi University (Milan) and the Academy of Management provided useful comments. The usual disclaimers apply.
HOW CAPABILITY DIFFERENCES, TRANSACTION COSTS AND SCALABILITY INTERACT TO SHAPE VERTICAL SCOPE

ABSTRACT

This paper provides an integrative analysis of the drivers of vertical scope, and formalizes theory on capability-driven integration, focusing on the following questions:

(1) How does the distribution of individual firm capabilities along a value chain affect vertical scope?
(2) How do transaction costs interact with capability distributions in vertically linked segments to determine scope at the firm and industry level?
(3) How do costs to growth in different parts of the value chain affect vertical scope?

To answer these questions, I use analytical and computational methods to build an evolutionary model where firms set their vertical scope in a competitive context. Specifically, I propose a model with two vertical segments (upstream and downstream), firm populations with heterogeneous capabilities, and an intermediate market subject to transaction costs, where firms can choose freely if they are integrated or vertically specialized. By varying the level of transaction costs and changing the structure of the correlation of upstream-downstream capabilities in the industry, as well as of the costs of expansion, I generate numerical results that explain how vertical integration evolves over time. My results suggest that without capability differences, even if transaction costs are nil, firms remain integrated; transaction cost reduction is only relevant if there exists heterogeneity of firms along the value chain, or if there are bottlenecks (uneven limits to growth or costs of expansion) in different parts of the value chain. The model also explains “mixed governance” (concurrent use of the market and of integration) often observed in practice.

Keywords: Vertical Scope; Capabilities; Transaction Costs; Scalability
Over the last thirty years, significant progress has been made in understanding what determines vertical scope (Williamson, 1975, 1985, 1999; Shelanski & Klein, 1995). Most of that work has focused on a particular, narrower line of inquiry, which takes the transaction as the unit of analysis and examines the particular transaction costs that, on the margin, push a firm towards vertically integrating as opposed to procuring through the market. However, several authors have recently argued that, in addition to transaction costs, capabilities also have to be taken into account in explaining vertical scope (Argyres, 1996; Poppo & Zenger, 1998; Schilling & Steensma, 2002; Madhok, 2002; Leiblein & Miller, 2003; Jacobides & Hitt, 2003). Despite this current 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), though, 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.

This paper provides such a concrete set of tools that expand traditional views on the determinants and dynamics of vertical scope. Drawing on Ricardo’s (1817) theory of comparative advantage, Coase’s (1937) initial framework and Penrose’s (1959) insights, I provide a stylized analysis of how productive capability differentials combine with transaction costs to affect the vertical scope at the firm and industry level. 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; I then consider how transaction costs catalyze scope. This model also explains how purely dynamic factors, such as the constraints to growth and expansion of firms in different parts of the value chain, affect vertical integration, and by doing so provides an explanation for the “mixed governance” (concurrent use of both vertical integration and market procurement) often observed in practice.

To achieve these results, I propose a behaviorally grounded, simple formal model, which departs from the existing mode of analysis. Rather than looking at each firm in isolation, and each transaction separately, as has been done to date, I jointly consider all firms making 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 instead. I build on recent advances in dynamic non-linear programming to develop such a model, which shifts the mode of analysis from the extant comparative statics “ceteris paribus” approach, to systemic analysis of firms that concurrently produce and set prices, quantities and scope in the context of competition. This shift in perspective is predicated on the belief that we cannot and should not analytically separate the analysis of the division of labor between different firms and the division of labor across the vertical divide, i.e. vertical specialization (Jacobides & Winter, 2003).
The paper starts with an overview of the literature on vertical scope, focusing on the growing
discussion on how capabilities affect scope. I then provide the theoretical and methodological
foundations of the model, before describing it and motivating its analytical structure. I then move
to the model’s results, and in particular discuss how capabilities and transaction costs co-determine
vertical scope, statically and dynamically, and how the limits and costs to expansion (i.e.,
scalability) determine the extent and evolution of specialization. The paper concludes with a
discussion, and with this model’s implication for theory and practice.

Transaction Costs, Capabilities and Scope: From Ceteris Paribus to a Systemic Analysis

How firms determine their vertical scope is a subject of long-standing interest. A branch of
microeconomics, and a large part of new institutional economics, has been focused on understanding
and explaining what affects the decisions of firms to integrate vertically (Perry, 1989; North, 1986;
Williamson, 1991). Much of the existing research is based on Transaction Cost Economics (TCE),
which argues that firms choose their scope as a function of the transaction costs they face: If there is
a risk of ex post opportunistic renegotiation in dealing with a vendor, because a long-lasting, 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 commit asset-specific
investments will refrain from such investments, to avoid being held up. 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 transaction costs, themselves the result
of underlying asset specificity and hence the potential for 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 existing research has been to demonstrate that TCE provides accurate predictions, in
that asset specificity does matter. Capability differences are rarely, if ever, considered. Riordan and
Williamson (1985), for instance, argue that even with inter-firm production cost differences,
transaction cost-based explanations hold sway. Yet the authors do not detail how the analysis of
capability differences and transaction costs operate in concert. Over the last few years, though, 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
has been recently 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.¹

The question, though, still remains: Beyond highlighting the role of capabilities and resources in the
determination of vertical integration, what can we establish about how resource and capability
differentials combine with transaction costs to affect the evolution of vertical scope? A first answer
could be that “firms specialize in what they are good at, transaction costs permitting.” This self-
evident proposition brings to the fore a truism, lost in much of the research in TCE: When a firm
considers whether it should specialize or not, it does not compare itself to the “market,” but to other
firms. Surprisingly, the discussion in the literature almost tends to neglect that “the market” is, in
essence, “a firm.” The “market” does not, in itself, produce anything. “The market” is an
organizational interface, behind which is another firm -- a firm that finds it advantageous to sell on
the basis of its own productive capabilities. This point was noted by Coase himself (1988: 38, also
quoted in Williamson, 1999), who was concerned about lack of attention to this very issue in the
existing literature:

“[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.”

Nominally, much of the TCE analysis (and orthodox economics) acknowledges that production
differentials might play a role. Yet in practice, capability differences are underemphasized. As
Langlois and Foss (1999) argue, new institutional economics, and its application to strategic
management, have focused on exchange conditions, to the detriment of the analysis of production
conditions (cf. Richardson, 1972). Demsetz also highlights the lack of attention in TCE to the
differences in productive capabilities between firms, and remarks that

“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.” (Demsetz, 1988: 147)

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

¹ On the empirical side, the literature on how capability differences drive scope is also very limited, despite some
puzzling early findings. 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 capabilities were a significant driver of vertical scope, in addition to transaction costs. The growth of the
capability- and resource-based view of the firm further led scholars to consider the role of firm heterogeneity: Poppo and
Zenger (1998) considered skill sets and scale as determinants of the decision to outsource IT services; Leiblein and
Miller (2003) found a strong correlation between insourcing the production of a specific component and past experience
producing the component. In a more direct test of the relative explanatory power of different theories, Jacobides and Hitt
(2003) found that the amount of variance in integration explained by capability differences is at least an order of
magnitude greater than that explained by transaction costs.
“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)

So if we consider the differences in the productive abilities of different firms, the relevant choice is between the firm, and its cost levels (if the input is produced) as opposed to the cost levels of another firm, including its profit margin, after deducting transaction costs.

While this appears to be a non-controversial statement, it has profound implications. It suggests that a firm cannot use “the market” to meet a particular need unless somebody is willing to take the other side of the transaction. That is, we need to consider what a firm’s productive capabilities (i.e., the efficiency with which it converts inputs into outputs) are as compared to other firms’ productive capabilities, and then consider how transaction costs, given different distributions of capabilities in the industry, affect vertical scope, at the firm and at the industry level. Drawing an analogy to the theory of international trade (Ricardo, 1817; Deardorff, 1980), I argue that to understand the patterns of specialization we have to examine the comparative advantage of all those who participate or could participate in the production process.

Since we have to consider the population of firms, and how it interacts competitively, we also have to change our mode of analysis. So far, transaction cost economics has focused on a partial equilibrium analysis—it has considered the conditions a firm faces as given, and has looked at how the determination of vertical scope happens, on the margin. Even the recent suggestions of Williamson (1999) and Madhok (2002) for including capabilities in the firm’s calculus of choosing its scope keep with this ceteris paribus view, suggesting we re-cast the choices of firms “on the margin.” However, as I point out above, such an approach may be misleading, since we have to consider all parties to the potential transaction. To do so, we need to shift from the partial to the general equilibrium. Another important reason for which a systemic analysis is called for is that this will enable us to consider the implications of competition and selection on vertical scope. Vertical scope is only one of the choice variables a firm has to make, and our analytical framework must accommodate 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 to choose whether and how much specialization is appropriate for them, and if they will focus on one or another segment, thus endogenously and incidentally determining vertical scope, as a function of both transaction costs and capability differences.

This suggests that even if our objective is to understand an individual firm’s decision in terms of vertical scope (i.e., “do I specialize or not?”), we need to have first understood what shapes the firms’ environment, which 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”). That is, to better understand decisions at the firm level, we have to study the industrial system as it evolves.

Finally, I argue that to better understand vertical scope, we have to take an explicitly dynamic framework. This would also enable us to shed some light on the studies of industry evolution, that have so far stayed clear of the discussion of how vertical scope evolves over time, with few exceptions (Langlois & Robertson, 1995; Christensen et al, 2002).

To provide a framework that explains how transaction costs and capabilities co-evolve, shaping the division of labor between firms and vertical specialization over time, our intuition may be somewhat limiting, as too many factors co-vary. For this reason, formal reasoning, partly inspired from a recent field-study (Jacobides, 2004) can be of assistance. Concretely, the proposed formal model and attendant numerical simulation explore the following questions:

1. How do differential capabilities and their distribution affect the choice of vertical scope?
2. How do transaction costs combine with capability distribution to determine integration?
3. How do dynamic constraints, such as limits to expansion (scalability) affect scope?

The model builds conceptually on dynamic programming and activity analysis, using a Mixed Complementarity approach (Ferris & Kanzow, 2002). It also includes a “Marshallian Dynamic” component, in that the factors that are fixed in the short term are altered by firms in the long term. In keeping with the “history-friendly” approach (Nelson & Winter, 1992; Malerba et al, 2003) particular attention has been paid to the model’s behavioral plausibility- within the bounds any formal analysis inescapably imposes. To illustrate that, the next section provides the intuition 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. Our aim is to explore what happens in an industry where firms have differing capabilities in different parts of the value chain, and then consider how capabilities combine with both transaction costs and expansion costs to drive scope. To do this, I construct 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.) 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

---

1 In terms of prior modeling literature, a notable yet little cited contribution is Rubin (1973). While Rubin is interested in formalizing Penrose’s (1959) insights on limits to expansion and apply them to M&As and corporate control, the model presented in the next section shares some analytical similarities to his: Both models use programming analysis, treat resources and capabilities as distinct elements, and posit that limits to expansion matter.
whether they consume their intermediate product, sell it to other firms, or buy it from other firms, and thus incidentally determine vertical scope. They also decide the prices at which they trade, both at the intermediate and at the final markets. In order for a firm to produce in each segment, it needs resources -- a segment-specific bundle, for which it pays a price. The supply function for both upstream and downstream resources is known and fixed in the beginning of the period; the price for resources is set endogenously as a function of the total demand by the firms that want to produce. Finally, firms are endowed with heterogeneous capabilities, both in terms of upstream / intermediate and downstream / final production. Capabilities are defined in the model as the way in which (i.e., the efficiency with which) resources are turned into products. Figure 1a illustrates the model’s basic structure.

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” for the resources they can use (the maximum number of branches, or limit on number of employees, etc.). Each firm knows its capabilities and transaction costs (on which I will elaborate in the next sub-section), and decides production of intermediate and final goods by deciding how much of the resources will be bought and used. It decides its production level, and may select not to fully utilize its capacity, if it is not profitable, as Figure 1b indicates. As a function of the total demand for resources, the price of the resources is set endogenously; likewise, as a result of the supplied quantities (at given prices) by the firms, and the elastic demand for the final good, the market clears and a price for the good obtains.

To provide an illustration, if this were the cement industry (cf. Ghemawat & Thomas, 2003), the 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 with which this resource bundle (which can be bought on the market, and whose price is a function of that resource bundles’ industry-wide demand) can be turned into cement. The downstream segment would be the cement mixing business, for which cement is needed in fixed proportions, and for which another bundle of resources (garble, 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 by the final customers, who are price-sensitive. Firms can trade in the intermediate good (cement) by selling or buying it at any quantities, in prices which 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).
Theoretically, then, our model considers the differences between productive capabilities, which are consistent with Winter’s (2003) recent discussion of “zero-level” capabilities, i.e. the relative efficiency with which firms turn inputs into outputs in each part of the production process. We expect that these productive capability differences, largely ignored in the literature, are quite important in practice (cf. Lieberman and Dawhan, 2001). Productive capabilities rest on the firm’s general and specific knowledge of how to do things (Richardson, 1972; Teece, Pisano and 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, 1991, Levinthal, 1997), even if all firms have access to homogeneous resources and compete in the market for these resources -- as they do in our model. To sum up, our model considers the impact of different types of distributions of productive capabilities along the value chain, and how these interact with transaction costs to determine vertical scope at each period.

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 its pattern of expansion or contraction per segment, and as such may also affect its capabilities in the next period. With these new Marshallian “fixed factors” in place, the new period begins, competition and production happen, and the cycle repeats itself.

To return to our cement example, the dynamic part of the model would be akin to the adjustment of the firms’ capacities, both in the upstream and the downstream segment. Firms, on the basis of how successful they were, would plan for their expansion, potentially paying cost if expansion is costly, i.e. if the productive capability is not scalable. This provides the new productive capacity / maximum resource endowment limits (the plants, maximum number of employees, etc.) with which the firm competes in the next period.

To conclude the model’s illustration with our industry example, the key questions the model examines is whether firms would specialize in cement production, in mixing, or would integrate by being active in both, as a function of the distribution of productive capabilities (efficiency of turning inputs into outputs) and of transaction costs in each period. Additionally, it considers how such specialization in production vs. cement mixing evolves over time, as a function of both individual firms’ choices on scope, and changes in the structure of the industry as a function of natural

---

1 Note that an advantage in terms of productive capability leads either to superior quality, or to inferior costs per output of unit. In the model, we consider a homogeneous good or service, with differences in productive capabilities being equated to differences in costs per unit of output. This is just a convenience, as any 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 given, quality-adjusted unit of output.
selection (e.g. if specialized firms grow more than integrated ones, pushing the industry to greater vertical specialization).

[1] The full model is included in the Appendix; I here briefly sketch out the intuition and the verbal description of the analytical structure, starting with the individual firm’s decisions in each period, then moving onto how firms interact through the final and intermediate markets, and conclude with the dynamic update 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 either intermediate or final goods; and to produce final goods it needs a given quantity of intermediate goods. To do so it can use its own intermediate good, buy from the market or do both; similarly, it can decide it produces more intermediate good than it uses up itself, and thus be 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 resources clears at some price, and this determines the patterns of trade, with firms either selling or buying resources from a central trade pool. (That is, the intermediate resource is not differentiated). Similarly, the final good is not differentiated and the total supply in that market is determined by the aggregation of the supply of the firms in the industry, whereas the total demand is determined by an elastic demand function.

The cost of the resources a 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 firms’ demands, and the supply some price elastic supply. 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 if they buy or sell intermediate good, and at what price, under given constraints. In addition, the global equilibrium conditions ensure that the intermediate and final goods and resources markets

4 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, ch. 6 and 7 for the seminal expositions), even though not much recent research has been written 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 done.

5 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 does provide some interesting results, but exploring them would make the exposition overly complicated. As such, the analysis of elasticity and differentiation has been left as an extension to the paper.
clear. The resulting problem belongs to a class of models known as Mixed Complementarity Problem (MCP), for which solution algorithms exist in the computational general equilibrium literature (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 transaction costs, maximum capacities, and capability levels (which are scale-insensitive). Solving their profit-maximization for both the upstream and the downstream divisions, they decide the level of production on both the final and the intermediate good, thus affecting its degree of integration (vs. use of the market) and total firm-level profitability.

Dynamic Update: Expansion of Endowment, 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. After the period is over, firms change the factors that were fixed in the previous period. Specifically, they first update their resource endowments. If they used up more than a certain percentage of their endowments, they expand; if they produced below capacity, they shrink to come closer to actual demand. If the firm expects that its optimal production quantity in the next period is roughly the same as in the current period, it is natural for the firm to require that its future maximum capacity be somewhat higher than the capacity it currently uses- we denote that with the “Capacity Buffer”, which means that each firm will want to have a maximum capacity which is a multiple of (i.e., percentage over) what capacity was used last 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 or take over market shares from their competitors. Even in declining demand scenarios, 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 the static model.) The “growth aim” of firms (i.e., the value of this capacity buffer) is a parameter, which is not varied in the model, since it does not change the qualitative results of the model (its value was set to 1.3); 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.

Expansion, though, is not always costless. When firms increase their capacity, they may face diseconomies of expansion. The intuition behind this penalty is that an efficient firm cannot hope to replicate its success exactly as it grows. This is consistent not only with views in strategic management or evolutionary economics on problems of replication in firms (Winter, 1995;
Szulanski, 1996; Winter and Szulanski, 1998), but also with more traditional views in economics on rising internal adjustment costs (Jorgenson, 1967; Lucas, 1967; Rubin, 1973). The extent of this cost of expansion is represented in the model by $\varepsilon^{\text{Exp}}$, the elasticity of expansion.

The elasticity of expansion varies by segment. In the results section, I first consider zero expansion costs, i.e. costless expansion, which means that $\varepsilon^{\text{IExp}} = \varepsilon^{\text{FExp}} = \varepsilon^{\text{Exp}} = 0$; non-zero but symmetrical expansion costs for both segments, which means that $\varepsilon^{\text{IExp}} = \varepsilon^{\text{FExp}} = \varepsilon^{\text{Exp}} > 0$; and finally expansion costs which are not identical for each segment, that is $\varepsilon^{\text{IExp}} \neq \varepsilon^{\text{FExp}}$.

Finally, another type of “adjustment limit” is introduced, in that no firm can increase or shrink its capacities by more than a factor $\Delta_{\text{Imax}}$ upstream and $\Delta_{\text{Fmax}}$ downstream. While in most of the analysis this constraint is not binding, I consider how the results change when it does become binding, and how this “temporary” limit to growth (motivated by transient limits to the speed of expansion) compares to downright costly expansion.6

This completes the discussion of how much firms want to expand in the intermediate resource market given that they are boundedly rational, and reasonably myopic.7 With the new fixed factors in place, the next period is run, output, prices, profitability and integration are determined, and 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 model, transaction costs represent the added costs of procuring through the market rather than producing in-house. These costs are akin to a “net tax,” levied each time the market is used. The firm that sells the intermediate good sells it at some price; the firm that buys it pays that price plus the transaction costs. This conceptualization, consistent with Masten, Meehan and Snyder (1991)

---

6 The two dynamic limitations we imposed work in different ways -- as they represent different types of dynamic problems, $\varepsilon^{\text{Exp}}$ represents the irreversible problems a firm faces in expanding beyond its knitting, or spreading its existing capability too thin. It is not a temporary adjustment cost; rather, it is the price that is paid for growing too fast (see Porter, 1995b, for a motivation of “penalty” for growing). The growth limit to expansion set by $\Delta_{\text{Imax}}$ and $\Delta_{\text{Fmax}}$, on the other hand, is a cap which is imposed by the inability to grow fast, or possibly by the risk aversion of firms which know they can grow more, but do not want to do so quickly and risk hurting profitability.

7 At this point, we should make an extended aside to the degree of myopia / foresight implied in this model. In our model, firms react to the results they got within-period rationally; the model’s structure does not make any heroic assumption on 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 it 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 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. A major extension of this model is to prepare a new version, with actors blessed with greater foresight and strategic behavior, and compare its results with the current, boundedly rational structure.
and Gibbons (1999), allows us to portray transaction costs as frictional costs, which are a net welfare loss for society. This “tax”, which represents the extent to which costs of market transacting are greater than those of internal, bureaucratic costs (Williamson, 1985: ch.6; Milgrom & Roberts, 1992) may be due to a variety of factors. For instance, for fears of ex post facto rent expropriation, market-based procurement may be riddled with Williamsonian under-investment in co-specialized assets. This would mean that an intermediate good or service bought from the market is less valuable than the good made in-house, because of the lack of requisite asset specificity; market-based production is more generic, and hence less valuable (represented here as the “inefficiency tax” in using the market). Alternatively, procuring through the market may entail risks, captured by a “tax” in terms of an imputed expected cost which is associated with market procurement.

The conceptualization of transaction costs tax is also consistent with the frictional Coasean costs associated with using the market (such as setting up purchasing divisions, searching the market etc.), and to the measurement costs between two stages of the production process, which are generally higher when done outside the boundaries of one firm (Barzel, 1982; Masten, 1991).\(^8\)

Finally, this “tax” is also consistent with non-TCE approaches. The proposed “tax,” which varies from zero to a high percentage of the value of the intermediate goods, reflects 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. The costs of the TC tax can, of course, be offset from the expected gains from trade due to production cost differences; indeed, it is this tradeoff between TC tax and productive capability differences that is at the heart of this model.

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 transaction costs, statically and dynamically. The focus in this paper, though, is narrower: It is to understand what determines vertical scope at the firm and the industry level, and how that changes over time. Therefore, the objective is to understand how different capability distributions, and different expansion costs or expansion limits along the value chain, interact with transaction costs to determine the scope of individual firms, and, as a result, the degree of vertical integration / specialization at the industry level. More concretely, I consider how

\(^8\) Note that this is a “net tax”- that is, a tax which represents the additional risks of market transaction when compared to coordinating 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 negative- that is, in-house production can be less effective (regardless of productive capabilities) as it is riddled with bureaucracy and politicizing inefficiencies that markets do not have (Williamson, 1985: Ch 6, Milgrom and Roberts 1992). I thus “normalized” TC to start at zero, both for expositional 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.
changes in TC affect the model’s vertical specialization; and then study how different capability distribution and expansion cost structures change 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 4c) that illustrate the model’s results, one axis ($z$) is vertical specialization at the industry level; the other axis ($y$) is time, showing how specialization evolves over 30 periods; and the third axis ($x$) represents transaction costs, which vary from 0 to 100% of the intermediate products’ cost, at ten steps of 10%. Thus, each figure describes vertical specialization over time (for 30 periods), for range of TC values, for a given distribution of capabilities and expansion costs.

Furthermore, each figure corresponds to a particular set of capability distribution and expansion costs, which is what I call a “scenario”. I consider nine such scenarios. Each scenario corresponds to a particular set of initial capability distributions (high / zero / negative correlation between individual firms’ upstream and downstream capabilities at $T=0$) and expansion costs (no expansion costs / expansion costs in both segments / expansion costs in one segment only). This provides $3 \times 3 = 9$ scenarios, which are summarized in Table 1, and correspond to figures 2a to 4c.

Summing up, this research design allows me to consider within each scenario how changes in TC affect vertical specialization over time. By comparing across scenarios, I can further assess (a) what is the impact of initial capability distribution and expansion cost conditions on vertical specialization (as each scenario represents a different initial capability distribution / expansion cost combination); and (b) how capability differences and expansion cost structures moderate the relationship between TC and specialization (as TC vary within each scenario, yet the same TC changes do not yield identical impacts in each of the different scenarios).

All nine scenarios have the same overall mean and initial variation in capabilities by segment; that is, if we consider the variance of 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 up- and downstream ($\sigma=0.1$). The capability mean is also the same in all scenarios, set to $\text{Avg}[aI(i)] = \text{Avg}[aF(i)] = 1$, wlog. However, the scenarios differ with regard to the correlation between individual firms’ capabilities in the up- vs. the downstream segment. In the “asymmetrical capability distribution” scenarios (scen. 1, 4 and 7), there is a negative correlation between the upstream and the downstream initial capability of the firms in the industry, so that a firm is good either upstream or downstream. Conversely, in the symmetrical capability distribution scenarios (scen. 3, 6 and 9).
there is a very high positive correlation between the initial capabilities of the two vertical segments - in other words, whichever firm is good upstream is good downstream as well. In the zero-correlation distribution scenarios (scen. 2, 5 and 8), there is no systematic pattern in who is efficient up- vs. downstream; some firms happen to be good on both segments, others on the one, etc., so that being good upstream is not correlated at all with being good downstream. The fact that in all nine scenarios we have the same overall aggregate capability variation within the industry makes them comparable, and allows us to focus on the implications of the degree of correlation between capabilities along the value chain.

These three core types of scenarios were run with zero expansion costs: that yielded scenarios 1-3, depicted in figures 2a-2c. I then considered positive expansion costs in both parts of the value chain (upstream and downstream), thus constructing scenarios 3-6, which are illustrated in figures 3a-3c. Finally, I considered expansion costs only in one part of the value chain (downstream), thus constructing scenarios 6-9, shown in figures 4a-4c. Table 1 summarizes the research design.

\[ \text{Include Tables 1 and 2 about here} \]

**Parametric Choices and Robustness Checks.**

Substantial effort was put into verifying that the model’s results hold over a large part of the solution space, and that they are robust to changes in the parameters of the model. To do so, I considered how the results of the base model would change, in qualitative terms, if I varied the parameters within reasonable ranges. As Table 2 suggests, none of the results was reversed through the wide ranges considered; additionally, the way in which a parameter affected the results, was economically interpretable. Table 2 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 \( b_i(i) = b_F(i) = 0.9 \). For the construction of the “strongly negatively correlated” scenarios, I set to the correlation coefficient between the upstream and downstream segment capabilities \( r = -0.99 \); for the “strongly positively correlated” scenario \( r = 0.99 \); and in the

---

5 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.

10 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’ capability. For all other elasticity of production values, in both the intermediate \([b(i)]\) 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.
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, were considered, and in the more than 67 different versions of the base model, the results discussed in the next section still held; that is, the qualitative results reported were found to be robust to parameter permutations of all sorts.

Finally, in all 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 is initially predisposed to vertical integration. The model then considers whether integration will remain, or if and when specialization will emerge instead. This modeling structure is motivated by Stigler’s (1951) description of the evolution of vertical integration, where the industry starts vertically integrated and then, given opportunities to do so, it dis-integrates. More importantly, such a starting point makes the results easy to follow graphically. Yet 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 discussed in the paper still held up.

**Generating Scenario Implementations through Monte Carlo Simulations.**

Having ensured the robustness of the model in terms of parametric choices, I did another robustness check: Through Monte Carlo simulations, I constructed 15 different implementations of each of the 9 scenarios – i.e., scenarios with the same parameters and capability correlations, but with different micro-structures in the capabilities of each firm. I then considered 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 each and all of the 15 implementations we 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. In figures 2a-4c, then, each point represents the average value of 15 runs.

Summing up, each of the 9 scenarios described in Table 1 is run 15 times (with different randomly generated implementations), and within each scenario I consider 11 different TC values, and run the model for 30 periods. This leads to 44,550 different MCP problems, each individually solved with GAMS. Then, the results from these 44,550 runs were averaged across the 15 implementations; they

---

11 Specifically, I tried the model with an initial unbalanced capacity endowment (75% downstream / 25% upstream for some firms; the inverse for other firms). I considered three such versions: 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 setup where there was, perhaps perversely, a negative correlation between capacity and capability in the initial period. For all the resulting 3x3 = 27 graphs, representing 297 runs of 30 periods, the qualitative results discussed in the paper still held, even though the graphical results were more difficult to follow. Thus, the choice of a vertically balanced initial capacity structure was chosen on the grounds of both being analytically plausible, and consistent with prior literature, and because it yielded a simpler illustration.
were consolidated by scenario; and finally converted to the nine graphs (2a to 4c) through a custom-made MS Excel database.

Model Results:

**Capability Distribution, Transaction Costs, and Scalability as they shape Scope**

*Vertical Integration as a Function of Capability Distribution and Transaction Costs.*

The first important finding is that the correlation of 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, measured as the fraction of the intermediate good which is traded through the market, over the total intermediate good produced. An index of 1 means full specialization (firms specialize either up or downstream); and 0 means full integration (no intermediate good is traded in the market).

In Figure 2a, capabilities along the industry’s value chain are not well correlated: some firms are weak upstream and strong downstream, and vice versa. Given such imbalances, vertical specialization does occur; furthermore, given such capability dispersion along the value chain, the reduction of transaction costs leads to increasing levels of dis-integration. Even with substantial transaction costs (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, transaction costs have to fall below a critical level, and that is the level whereby the “tax” that transaction costs impose is offset by the benefits from accessing superior capabilities. In general, the threshold in terms of transaction costs where specialization starts emerging is a function of the magnitude of the capability differences between firms; given extreme differences in capabilities, bigger than the ones reported here, specialization occurs even with very high levels of transaction costs.12

In Figure 2b, we have zero correlation between upstream and downstream capability. A comparison with Figure 2a shows that both level and rate of increase of specialization are lower. Furthermore, the critical level of transaction costs below which specialization occurs is substantially lower than in Figure 2a. All these effects can be explained by smaller (average) capability dispersion within a

---

12 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 TC reduction will always encourage more specialization, except from the degenerate case of full (100%) specialization, where a TC reduction is no longer able to lead to more specialization.
firm, the attendant lack of opportunity for co-specialization between any two firms in the industry, and thereby the reduced motivation to trade.\textsuperscript{13}

In Figure 2c, by contrast, 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 gains to trade to be had by specializing. So even for low transaction costs, no specialization \textit{at all} occurs, as there isn’t any strategic logic for it.

By comparing Figures 2a, 2b and 2c, we observe 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 capability differences, and in particular the correlation of firms’ capabilities along the different value chain segments. It also obtains that transaction cost reductions \textit{do} play a role, but only when there is an underlying heterogeneity of capabilities along the value chain; in other words, transaction costs \textit{catalyze} scope through the underlying differences in capabilities. Much like in the theory of international trade, comparative advantage is what drives the patterns of trade, comparative advantage drives the patterns of vertical specialization. And much like international tax and transportation costs operate on the underlying productivity differences to shape international specialization (Ricardo, 1817; Deardorff, 1980), transaction costs 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. That is obvious from the upward slope of the specialization plane over time, and in our firm-level data.

The second interesting observation is that this dynamic effect of increase in specialization is more pronounced in lower TC levels; that is, low TC not only allow for a higher specialization in each period, but also enable greater \textit{rates of increase} of specialization over time. The reason is that lower TC allow the firms which are strong either downstream or upstream to expand more (for higher TC, such an expansion would have been less profitable, as it requires paying more of the TC “tax”). By doing so, the mix of production is increasingly tilted towards the specialized firms, and such specialization increases more quickly over time as the integrated firms in the industry shrink their production, and specialized producers, who are more effective, replace them. Figure 2b in particular

\textsuperscript{13} 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, strictly put, a \textit{very} low capability) in another. So the existence of such vertical specialists is consistent with the model’s setup.
shows that over time, there is a significant degree of specialization. This happens 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 capabilities, then even small transaction costs will be sufficient to deter them from using the market. On the other hand, if this is not the case – that is, if firms have widely varying capabilities, and in particular if their capabilities are uncorrelated or negatively correlated along the value chain, then even high transaction costs are not sufficient to offset the economic desirability from using the market, i.e., from having a vertically specialized industry structure. This brings us to the second major conclusion from the model -- and perhaps the most important one. Specifically, a reduction in Transaction Costs is the catalyst, but not the ultimate driver, of vertical dis-integration. (cf. Demsetz, 1988). If we compare Figures 2a, 2b and 2c, we observe that a decrease of transaction costs in the presence of asymmetrical capability distribution (where upstream and downstream capabilities are negatively correlated) will lead to significant specialization; yet if capabilities are symmetrically distributed, then a TC reduction will not produce any effect. In other words, a TC reduction lubricates the workings of the market and allows firms to capitalize on their capabilities and relative strengths. If firms look different, transaction cost reduction will allow for significant specialization, as each firm will focus on its area of strength; yet if all firms look alike, the same reduction will not do much (if anything) to promote specialization and dis-integration. No latent gains from trade will exist to change vertical scope. The insight, then, is that capability distributions (in particular, the correlation of 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 “symmetrical 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 in these cases specialization does not occur, even though some firms have an 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; not 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 of vertical scope that are determined by simple *comparative* advantage.

The dynamic relationship between transaction costs and capability distribution also makes economic sense. As industries evolve, the *rate of increasing specialization is partly a result of the pure forces of selection*. That is, increasing specialization over time in Figures 2a and 2b are *not* due to any exogenous change in TC; rather, it is due to the very fact that specialized firms take an increasing share of the production, as a result of being more competitive. The relative speed of increase of specialization is *significantly greater* for low TC values. This happens because lower TC enable a greater efficiency in the “hand of natural selection,” which means that lower TC can 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 of 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.

**Scalability, Limits to Expansion and Vertical Scope**

Scalability –the extent to which firms can grow, and do so without costs in growing -- also affects vertical integration. As discussed, the model considers two types of scalability issues. The first one is the cost associated with growth -- a sort of “growth costs” or “diseconomies” (cf. Rubin, 1973). The second is the existence of some “limits to growth,” that is, a cap in the rate of growth. In this section, I consider how the “cost of growth” affects scope, and provide the corresponding graphs; I then consider how “limits to growth” affect scope; and see how these two differ.

First, costly expansion (a high level of $\varepsilon^{\text{Exp}}$) dynamically tapers specialization. Specialized firms, as they grow, may erode their respective competitive advantage, and as such, over time they have less of a competitive edge to offer to other potential transactors. So over time, the underlying gains from trade are reduced, in relative terms, and as a consequence, specialization is more limited than would otherwise be the case. This can be seen by comparing Figure 2a to 3a and Figure 2b to 3b; the corresponding figures are identical, in all aspects other than the expansion dis-economy for both the upstream and downstream segments. Note that, as expected, Figure 3c is identical to 2c: Given capabilities that are highly positively correlated up- and downstream, no specialization takes place, either statically or dynamically.

The more interesting case, though, is the case of varied per-segment expansion costs. If each vertical segment has a different cost to expand, so that, for example, the upstream segment is easy to expand
whereas the downstream is not, then some intriguing results obtain. These results are most visible in
the low-capability dispersion case, where with symmetric expansion costs (be they positive or zero),
no specialization occurred, as we saw in Figures 2c and 3c. Figure 4c shows the same scenario, with
a substantial expansion cost downstream and no expansion cost upstream. Contrarily to the results in
2c and 3c, Figure 4c shows that substantial vertical specialization can emerge even in the absence of
capability heterogeneity along the value chain. So the unevenness of the expansion costs along the
value chain can, in and of itself, cause vertical specialization.

This result, even if unexpected, has an intuitive, and empirically plausible explanation: Efficient
firms will tend to expand in the segments in which they are efficient. If they are equally efficient in
both segments (as in the scenario of figures 2c, 3c and 4c), they will try to expand both their
segments; the most efficient firms will expand more aggressively, and the less efficient but good
firms a little less aggressively. The inefficient ones will shrink correspondingly, and ultimately
wither. When the expansion costs do not differ between the two parts of the value chain (figures 2c,
3c), firms will expand (or shrink) in both segments, and remain integrated. Yet if they can only
expand costlessly in one segment, and the other segments’ expansion is costly (figure 4c), then the
efficient firms will expand only in the segment they can expand, and for the one they cannot expand
they will rely on the other firms, which will find it profitable not to shrink both segments, but shrink
only the one that is more readily scalable (and as such has new competition) whereas they will
remain in these segments which are hard to grow.14 So if an efficient firm can only grow upstream, it
will expand upstream and will rely on another, less efficient firm for the downstream good.

Similar results obtain for the “limits to growth” case. The main difference is that in terms of
specialization, limits to growth that are identical to both segments of the value chain only work to
delay, rather than dynamically taper, specialization. In the case of limits to growth that apply only to
one of the two segments, specialization does result even if capabilities are symmetrically distributed,

14 This provides an interesting and unexpected modeling insight -- namely, that vertical specialization and Mergers &
Acquisitions 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 with organic growth, which is subject to the same limitations that organic growth in practice often is. It
cannot just go and buy the resource endowment of other firms in the industry. Given that this is the 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 the 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 the resources needed for downstream growth, it would possibly do that as a substitute to engaging
in specialization. This seems to be happening in banks: Inasmuch as the smaller players who have local distribution do
not want to be purchased, or inasmuch as there are regulatory barriers in such purchases (such as in Europe), then large
financial institutions specialize and use these smaller players in addition to their own branches to distribute their
financial products; the reason is that they can scale up the products (financial services) without much hardship, yet
cannot do so for distribution. In the US, in contrast, where smaller players with distribution presence are both more
interested in being acquired and are easier to acquire, rather than specialization, M&A takes place -- large financial
institutions satisfy their need for distribution through the purchase of branches. This does not happen mainly to avoid
transaction costs, but rather to enhance profitability through effective expansion, and transaction costs may be a
secondary driving force, as opposed to a primary one.
much like the case of the non-identical growth costs. There is, however, a significant difference: given sufficient time, such specialization recedes, as the efficient firms manage to catch up. This provides the interesting pattern of an endogenous cycle of specialization giving way to integration, as efficient firms initially need and use co-specialized firms, only to grow and abandon their transactional partners, when they can integrate, later.15

Overall, the analysis of the constraints to growth (i.e., scalability) suggests that vertical specialization can emerge as a result of the dynamic process of competition. With symmetric expansion costs or limits to growth, inefficient firms are simply weeded out as the efficient firms grow and compete them away; whereas with unbalanced expansion costs along the value chain, efficient firms “compete away” only the part of the value chain where they can grow. Vertical specialization thus emerges as an incidental result of these dynamic processes. With asymmetric capabilities, expansion costs may dynamically taper specialization.

Discussion

This paper set out to tackle three tasks through a formal model: First, analyze how capability distributions affect integration; second, examine how transaction costs and capability differences interact to shape scope; third, examine how dynamic factors such as scalability affect scope and its evolution. A number of insights obtained from this model. I briefly consider them, and suggest how they allow us to revisit extant theory.

First, I suggest that rather than consider TC and capabilities as additive factors, which all have to be taken into account by an individual firm, we have to look at the distribution of 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. Second, the analysis also shows how transaction costs operate on this distribution of capabilities, shaping vertical scope. A way to visualize this is to consider a lake, and think of the level of water in the lake as being the transaction costs. As transaction costs recede, if the landscape is rugged -- that is, if the capabilities are asymmetrically distributed -- islands of specialization will increasingly appear. Yet if the underlying surface is smooth, if there is no heterogeneity to uncover underneath the surface, nothing will change in the landscape, and integration will remain.

15 Note that in the case of asymmetric capability distribution, having a cost only downstream (as opposed to in both sectors), such as that in figure 4a ends up dynamically limiting as opposed to enhancing the extent of specialization, when compared to the zero-expansion penalty case (figure 2a) although specialization is still higher than it would have been if both segments had expansion costs (see figure 3a). The reason is that given asymmetric capabilities, the equally asymmetric cost equalizes the capabilities dynamically; firms, as time unfolds, look increasingly similar as the firms which were strong downstream pay “a price” to grow, and as such become more similar to any other, and as such the gains from trade become more limited. Therefore, the impact of the differential limit or cost to expansion depends on the initial capability distribution.
These conclusions have notable implications, as they refine our view on what really affects vertical integration. Economic theory (including TCE) abstracts away from productive capability differences and proceeds to examine what happens if they are trivial. The problem, though, is that this analytical convenience, necessary to derive closed-form formal results, has inadvertently turned almost into an assumption of empirical fact. This may lead to inappropriate predictions, especially as capability differences between firms appear to be very considerable (Lieberman & Dhawan, 2001). Once we bring capability differences back in, it becomes evident that capability differentials and the possibility of gains from trade are a key factor; transaction costs are also re-cast as the catalysts, but not the necessary and sufficient drivers of vertical scope.

Additionally, 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 occasionally be misplaced. In particular, that focus may make us overlook some economically important drivers of vertical scope, which depend in the comparative advantage and the distribution of productive capabilities, as opposed to competitive (absolute) advantages. Indeed, there is little, if any literature that explicitly considers how capabilities are distributed within a segment, and how they are correlated between segments; explicitly studying these issues holds much promise.

In terms of analytical structure, this model also represents a shift from the ceteris paribus analysis of the transaction to the analysis of the system of capabilities and transaction costs as it co-evolves. The relationship between such a systemic analysis of capabilities and transaction costs as they determine scope and division of labor and the micro-analytic inquiry of TCE is broadly analogous to the relationship between a general equilibrium inquiry acknowledging the causal interdependence of prices in different markets and the study of price determination in a single market. The advantages of addressing the narrower problem, in terms of feasibility and focus are enormous, but the broader problem provides a more satisfactory account of the causal logic (Jacobides & Winter, 2003). For instance, our model shows that every marginal “make-vs-buy” decision an individual firms makes, 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 capabilities of all participating firms (not any two firms), in both the upstream and the downstream segment. It also shows how these capability conditions, and thus prices, 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.
Framing the problem of vertical scope as one determined by industry-wide competitive dynamics, which depend on capabilities and TC, also has practical implications. In particular, it explains some theoretical puzzles, e.g. why TC reductions do not always lead to less integrated firms. The reason may be that TC operate on capability distributions, and that capability distributions differ between settings, industries, or countries. Another implication comes 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 “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 literature. It does so by combining the insights of Coase (1937) with the analysis of dynamic adjustment costs pioneered by Penrose (1959).

Specifically, the model predicts that mixed vertical strategies will be pervasive in the presence of differential scalability along an industry’s value chain. Qualitative evidence in a number of sectors I have been involved in corroborates this finding. Banks, for instance, try to sell their banking products both through their own branches, and through other intermediaries. The motivation for this, as I have been 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. For that reason, they use all of their internal production, and, when it becomes for them too costly to expand their retail branches, they also use “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. It is affected by history, not only because of reputation or path-dependency in contracting (Argyres and Liebeskind, 1999). Rather, we observe that the initial capability distribution and the adjustment of the resource and capability endowments significantly affect the degree of integration. This explains why vertical structure in similar industries may differ markedly between locales, markets, countries, or time periods; when the patterns of distribution of capabilities up- vs. downstream vary, vertical specialization also varies as a result. Thus differences in scope should not lead us to infer differences in TC; capabilities may well be the culprit. The model also suggests that the evolution of vertical scope itself is path-dependent, in that selection forces may well affect the dynamics of a sector as it evolves, through the virtual circle of specialization, reinforced by natural selection, as discussed in the previous section.

Another advantage of this model, which is explored in a follow-on paper due to space constraints, is that it yields specific predictions on the profitability impacts of changing transaction costs under

---

16 Such “mixed strategies,” pervasive in a number of sectors, have not been satisfactorily explained to date. While some arguments have been made in favor 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 presume 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).
different capability distributions, and as such allows us to consider how transaction costs 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 Jacobides, Winter, Georgakopoulos & Kassberger, 2004). This not only allows us to connect the analysis of vertical scope to business strategy, but also provides greater confidence in the results of the model, which are based on sound economic intuition.

**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 fully realistic or complete depiction of reality. Other limitations regard the model itself, the most important one being that the model is solved with 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 models’ analytical structure. Yet that being said, some of the particular findings of the model could be captured and supplemented by analytical proofs.

Furthermore, there are also limitations within the scope of modeling using computational techniques, and I intend to address some of these in future research. The objective is 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 am thus working on introducing transaction costs that increase as the number of suppliers diminishes; allowing for oligopolistic behavior when consolidation occurs; introducing endogenous margin-setting; allowing for strategizing in the markets for resources; modeling for heterogeneity in resources; considering multiple types of intermediate goods, with and without TC; etc.

Another limitation of the model is that it considers TC and capabilities as distinct categories. While this helps yield some tractable results, future modeling should allow for TC choices to directly affect productive capabilities. We should also consider how capabilities themselves are shaped by TC and the resulting vertical scope over time (Jacobides & Winter, 2003). Additional theoretical and modeling improvements should allow for capabilities to be endogenously changed by firms’ conscious efforts to do so, perhaps as a result of the aspiration level and attendant returns. This model could also be empirically tested, e.g. by measuring capability differences and considering how they interact with TC, or by explaining mixed governance choices.
These limitations aside, taking the “broader picture,” the potential contribution of this paper to the literature, and to the resulting prescriptions for managerial action, is the combination of three sets of considerations. The first of these is transaction costs. The paper acknowledges the existing conception of transaction costs, and expands their scope, by integrating the broader problems of communicating and coordinating across firm boundaries. The second set of considerations examined in the paper are capability differences. Taking firm heterogeneity seriously, the model provides explicit predictions on how capability differentials and growth rate constraints affect vertical scope - and profitability distribution. The final set of considerations is the analysis of purely dynamic elements such as the constraints to growth, and the issues of scalability, whose impact has not been considered to date, and which may well be responsible for the “mixed vertical strategies” combining integration and market procurement, often observed in practice.

The potential contribution of this paper is not so much the elaboration upon each of these three categories separately -- although significant effort is expended to that aim. Rather, it is bringing these three considerations together, to form a consistent body of theory. An example is the explanation of mixed modes of procurement, by combining the insights of Coase (1937) on the costs of market exchange with that of Penrose (1959) on dynamic adjustment costs. This analysis – and the model structure itself -- is also informed by qualitative and quantitative evidence alike, and lends itself to further empirical testing. The resulting integrative theory and analytical approach also addresses, through the same “toolkit,” both the question of determining vertical scope and the strategic ramifications of changing scope.

More important, the paper 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 expansion limits. 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 have to focus on issues that have not attracted attention so far, such as the distribution of capabilities across an industry’s value chain, and comparative advantage, or the limits to growth in different parts of the industry. Indeed, the contribution of this paper is not its technique, but rather its implications for framing and theory.

Last, the model, which has been partly inspired by recent qualitative fieldwork (Jacobides, 2004), 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 map the implications of changes of transaction costs 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 that contribute to its aggregate profit, which it aims to maximize. This leads to the following optimization problem faced by every firm $i$:

$$
\max_{(R_I(i), Q_I(i), Q_S(i), R_F(i))} \; Q_F(i) \cdot PF + Q_S(i) \cdot P_S - Q_IB(i) \cdot P_IB - R_I(i) \cdot P_RI - R_F(i) \cdot P_RF \\
\text{s.t.} \; Q_I(i) = a_{I(i)} \cdot R_I(i)^{b_{I(i)}} \\
Q_F(i) = a_{F(i)} \cdot R_F(i)^{b_{F(i)}} \\
Q_F(i) = \phi \cdot (Q_I(i) + Q_IB(i) - Q_S(i)) \\
Q_S(i) \leq Q_I(i) + Q_IB(i) \\
R_I(i) \leq R_{I\text{max}}(i) \\
R_F(i) \leq R_{F\text{max}}(i) \\
R_I(i), R_F(i), Q_S(i), Q_IB(i) \geq 0 
$$

where:
- $Q_I(i)$ is the quantity of the intermediate product that firm $i$ wishes to supply to the intermediate market for any given price
- $Q_IB(i)$ is the quantity that a firm demands from the intermediate goods markets for 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) that clears the intermediate goods market with an industry-level price for intermediate goods
- $a_{I(i)}$ is the capability of intermediate production function of firm $i$ (updated each period)
- $a_{F(i)}$ is the capability of final production function of firm $i$ (updated each period)
- $b_{I(i)}$ is the exponent for the intermediate production function of firm $i$ (constant over time)
- $b_{F(i)}$ is the exponent for the final production function of firm $i$ (constant over time)
- $\phi$ is the Leontieff production coefficient
- $R_I(i)$ is the quantity of intermediate resource used in production process
- $R_F(i)$ is the quantity of final resource used in production process
- $Q_I(i)$ is the quantity of intermediate product the firm produces
- $Q_F(i)$ is the quantity of final product that firm $i$ wishes to supply to the final market for any given price
- $P_F$ is the price of the final product
- $P_RI$ is the price at which the firm buys the intermediate resource
- $P_RF$ is the price at which the firm buys the final resource
- $P_IB$ is the price at which the firm buys the intermediate good
- $P_S$ is the price at which the firm sells the intermediate good
- $R_I(i)$ is the quantity of the intermediate resource used
- $R_F(i)$ is the quantity of the final resource used
- $R_{I\text{max}}(i)$ is the available intermediate resource endowment (also called maximum capacity)
- $R_{F\text{max}}(i)$ is 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 $Q_I(i)$ units of intermediate output, through the relationship:

$$
Q_I(i) = a_{I(i)} \cdot R_I(i)^{b_{I(i)}}, \; a_{I(i)}>0, \; b_{I(i)}>0.
$$
with \( b(i) \) being the elasticity of production for the intermediate good. Moreover, total production cannot exceed the limits posed by the available resource endowment,

\[
(1b) \quad R_I(i) \leq R_{I_{\text{max}}}(i).
\]

The firm-specific capability \( a(i) \) 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 have to ensure that the quantity of intermediate sold, is either produced or bought\(^{17}\):

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

For final goods, the firm-specific production function describes the capability of turning \( R_F(i) \) units of final resource into \( Q_F(i) \) units of output, through the relationship\(^{18}\):

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

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

\[
(2b) \quad R_F(i) \leq R_{F_{\text{max}}}(i)
\]

As in the intermediate good production, \( a_F(i) \) and \( R_{F_{\text{max}}}(i) \) are given and fixed within 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:

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

With \( \phi \) being the Leontief 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 \( R_I(i), Q_{\text{IB}}(i), Q_{\text{IS}}(i), R_F(i) \) (which also determine \( Q_{\text{IP}}(i) \) 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 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 \( P_{RI} \) and \( P_{RF} \); for intermediate goods bought or sold, \( P_{IS} \) and \( P_{IB} \); and for the final good, \( P_{F} \) ) 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 up to now 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

---

\(^{17}\) 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 acting as a broker, buying from other firms their intermediate production, and selling it to yet different firms; in equilibrium, of course, such arbitrage will not happen.

\(^{18}\) For both final and intermediate goods, we can generalize the formulation to the multidimensional case of vectors of inputs which are transformed into output, and as such consider the resource price as the price of the resource vector.
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 our model to equilibrate and a solution to occur, 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 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.19

\[(\text{I}) \quad S(PRI) = SI \cdot PRI^{\varepsilon_{SI}} \geq \sum_i R(i) \rightarrow PRI\]

Second, the market for final resources has to clear through PRF. Total demand is obtained by adding up individual demands that results 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:

\[(\text{II}) \quad S(PRF) = SF \cdot PRF^{\varepsilon_{SF}} \geq \sum_i RF(i) \rightarrow PRF\]

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:

\[(\text{III}) \quad \sum_i QIS(i) = \sum_i QIB(i) \rightarrow PIS\]

\[PIB = PIS \cdot (1 + TC)\]

Transaction costs are thus considered a net outflow from the system; they are a tax that represents friction. This conceptualization is discussed in greater detail later in 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 one for the final product. The supply for 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:

\[(\text{IV}) \quad D(PF) = DF \cdot PF^{\varepsilon_F} \geq \sum_i QF(i) \rightarrow PF\]

Parameter $\varepsilon_F \geq 0$ is the elasticity of demand for the final good, PF is the price of the final good, and

19 The arrow indicates the price, which corresponds to the market clearing in this general equilibrium equation.
parameter DF is a demand constant. This ensures, that, in equilibrium, demand is equal to supply (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.20 Therein, firms select how much they produce up- and down-stream, as well as if 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, 1997, for applications).

The MCP structure allows us to solve concurrently these optimizations, 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 Capabilities*

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

\[(A1a) \quad (1+\text{CapResI}(i)) \cdot Rl(i) \text{ and} \]
\[(A1b) \quad (1+\text{CapResF}(i)) \cdot RF(i).\]

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.

In addition, as discussed in the text, expansion is not always costless. In order to model these expansion dis-economies, capabilities \(aI(i,t)\) and \(aF(i,t)\) of a firm in turning input into output are expected to deteriorate as output grows, with given elasticities of expansion, \(\varepsilon_{IExp}, \varepsilon_{FExp} \geq 0\). More precisely, the functional form selected proposes that:

\[(A2a) \quad aI(i,T+1)=aI(i,T)\min\left\{1,\left(\frac{Rl_{max}(i,T)}{Rl_{max}(i,T+1)}\right)^{\varepsilon_{IExp}}\right\} \text{ and} \]

---

20 In order to confirm that our results are correct, we also solve a different 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.
That is, a firm’s capability deteriorates as it expands, and when it contracts to its original production levels, it can regain its capability; but its capability will not keep growing if it further contracts its production. The insight is that expansion may be costly, but the costs are reversible; yet firms cannot hope to become more efficient than they originally were just by scaling down.

For each of the two activities, the expansion penalty has been set to be identical for some of the modeling analysis, that is, $\epsilon_{IE} = \epsilon_{FE} = \epsilon_{F}$. However, the expansion penalties do not have to be identical for each segment so that in principle $\epsilon_{IE} \neq \epsilon_{FE}$.

Finally, another type of “adjustment limit” is introduced, in that no firm can increase or shrink its capacities by more than a factor $\Delta_{I}$max upstream and $\Delta_{F}$max downstream. Each firm, then, decides its expansion policy, and the following two update rules apply:

\[(A2c)\quad R_{I\max}(i) := \max\left\{ \min\{ (1+\text{CapRes}_{I}(i)) \cdot R_{I}(i), (1+\Delta_{I\max}(i)) \cdot R_{I\max}(i) \}, \right.\]
\[\left. (1-\Delta_{I\max}(i)) \cdot R_{I\max}(i) \right\} \]

\[(A2d)\quad R_{F\max}(i) := \max\left\{ \min\{ (1+\text{CapRes}_{F}(i)) \cdot R_{F}(i), (1+\Delta_{F\max}(i)) \cdot R_{F\max}(i) \}, \right.\]
\[\left. (1-\Delta_{F\max}(i)) \cdot R_{F\max}(i) \right\} \]

where
\begin{align*}
\Delta_{I\max}(i) & \quad \text{is the maximum relative change per period for intermediate capacity} \\
\Delta_{F\max}(i) & \quad \text{is the maximum relative change per period for final capacity} \\
\text{CapRes}_{I}(i) & \quad \text{is the relative capacity-reserve over current production } R_{I}(i) \text{ the firm aims at} \\
& \quad \text{for its intermediate division} \\
\text{CapRes}_{F}(i) & \quad \text{is the relative capacity-reserve over current production } R_{F}(i) \text{ the firm aims at} \\
& \quad \text{for its final division.}
\end{align*}

\textit{Linking Statics and Dynamics}

The final bit 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 capabilities can be updated using (A1). Furthermore, on the industry level, demand and resource supply are updated according to

\[(A3a)\quad S_{I}(T+1) = S_{I}(T) \cdot (1 + \Delta S_{I}) \]
\[(A3b)\quad S_{F}(T+1) = S_{F}(T) \cdot (1 + \Delta S_{F}) \]
\[(A3c)\quad D_{F}(T+1) = D_{F}(T) \cdot (1 + \Delta D_{F}), \quad \text{with} \]
\begin{align*}
\Delta S_{I} & \quad \text{the relative change in } S_{I} \text{ per period} \\
\Delta S_{F} & \quad \text{the relative change in } S_{F} \text{ per period} \\
\Delta D_{F} & \quad \text{the relative change in } D_{F} \text{ per period.}
\end{align*}

On the basis of these updates, the static model is run for the new period $T+1$, and the cycle starts anew; statics are linked to dynamics, and we may run an arbitrary number of periods.
References


Table 1: Research Design and Resulting Scenario Structure

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Upstream / Downstream Capability Correlation</th>
<th>Expansion Costs</th>
<th>Transaction Costs</th>
<th>Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Negative Correlation (r = -0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>2a</td>
</tr>
<tr>
<td>2</td>
<td>Uncorrelated (r = 0)</td>
<td>None</td>
<td>0-100%</td>
<td>2b</td>
</tr>
<tr>
<td>3</td>
<td>Positive Correlation (r = 0.99)</td>
<td>None</td>
<td>0-100%</td>
<td>2c</td>
</tr>
<tr>
<td>4</td>
<td>Negative Correlation (r = -0.99)</td>
<td>In both segments</td>
<td>0-100%</td>
<td>3a</td>
</tr>
<tr>
<td>5</td>
<td>Uncorrelated (r = 0)</td>
<td>In both segments</td>
<td>0-100%</td>
<td>3b</td>
</tr>
<tr>
<td>6</td>
<td>Positive Correlation (r = 0.99)</td>
<td>In both segments</td>
<td>0-100%</td>
<td>3c</td>
</tr>
<tr>
<td>7</td>
<td>Negative Correlation (r = -0.99)</td>
<td>Only downstream</td>
<td>0-100%</td>
<td>4a</td>
</tr>
<tr>
<td>8</td>
<td>Uncorrelated (r = 0)</td>
<td>Only downstream</td>
<td>0-100%</td>
<td>4b</td>
</tr>
<tr>
<td>9</td>
<td>Positive Correlation (r = 0.99)</td>
<td>Only downstream</td>
<td>0-100%</td>
<td>4c</td>
</tr>
</tbody>
</table>

Table 2: 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 adjust-ment swings; lower values partly taper specialization cf. footnote 13</td>
</tr>
<tr>
<td>Elasticity of expansion</td>
<td>0 or 0.4</td>
<td>0 to 1.5</td>
<td>Not reversed</td>
<td>Higher coeff =&gt; bigger impact</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>Capability variation per (within) segment</td>
<td>$\sigma = 0.1$ $\sigma^2 = 0.01$</td>
<td>$\sigma : 0.05$ to $0.2$ $\sigma^2 : 0.0025$ to $0.04$</td>
<td>Not reversed</td>
<td>Ceteris Paribus, higher variation (in $\sigma$) =&gt; higher specialization</td>
</tr>
<tr>
<td>Capability correlation up- vs. down-stream</td>
<td>-0.99 or 0 or 0.99</td>
<td>-1 to +1</td>
<td>Not reversed</td>
<td>Three near-extreme correlation structures chosen for illustration Other values yield same insight</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>
</tbody>
</table>

NB: For each scenario discussed in Table 1, and with the parameter values of Table 2, I constructed through Monte Carlo simulations 15 different versions of the scenarios (with different capability values for the 21 firms). All the resulting graphs (9x15 = 135 graphs, representing 1,485 runs of the model, i.e. 44,550 MCP solutions) confirmed the qualitative discussion in the paper. Figures 2a to 4c thus report the averages of each of the 15 different specific versions of the 9 scenarios constructed with Monte Carlo.
Figure 1a:  
(Static) Model Overview: The Basic Structure

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

- **Firms:** Maximize Joint Profits
  - Integration is an incidental result of profit-maximization

- **Upstream Production**
  - **Fixed Firm-Level Variables:**
    - 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
    - Intermediate Quantity Bought from other firms
    - Intermediate Quantity Sold to other firms
    - Intermediate Quantity Transferred Downstream

- **Downstream Production**
  - **Firm-Level Variables:**
    - 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
Figure 1b:
The Short-term (Static) Model: A Mixed Complementarity Solution

Intermediate market trade pool

Firm $i$

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

Upstream Resource Price

**Downstream Division**
*Maximize Profit*, by
- Selling to final market
subject to
- Capacity Constraint (resource endowment)
- Production Function
- Needed upstream input
Given:
- Downstream Resource Price

**Firm $j$**

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

**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

$Q_{IS}(i)$

$Q_{IB}(i)$

$Q_{IB}(j)$

$Q_{IS}(j)$

Intermediate Good Price: $P_{IB} = P_{IS}(1+TC)$

$\sum_{i} Q_{IS}(i) = \sum_{i} Q_{IB}(i)$ for all firms $i, j, ...$

Total Supply for final good:

$Q_{F} = \sum_{i} Q_{F}(i)$

Demand for final good:

$DF \cdot PF^{eF}$

Total Demand for final resource:

$\sum_{i} RF(i)$

Supply for final resource:

$SF \cdot PRF^{eSF}$

Total Demand for intermediate resource:

$\sum_{i} RI(i)$

Supply for intermediate resource:

$SI \cdot PRI^{eSI}$

$Q_{F}(j)$

$Q_{F}(i)$

$Q_{F}(j)$

$Q_{F}(i)$

$RF(i)$

$RI(i)$

$P_{F}$

$PRF$

$PRI$
Figure 2a:
Aggregate Specialization & Transaction Costs – Negative Up / Downstream Capability Correlation

Figure 2b:
Aggregate Specialization & Transaction Costs – Zero Up / Downstream Capability Correlation

Figure 2c:
Aggregate Specialization & Transaction Costs – Positive Up / Downstream Capability Correlation
Figure 3a: Aggregate Specialization & Transaction Costs –
Negative Up / Downstream Capability Correlation, with Expansion Costs (in both Segments)

Figure 3b: Aggregate Specialization & Transaction Costs –
Zero Up / Downstream Capability Correlation, with Expansion Costs (in both Segments)

Figure 3c: Aggregate Specialization & Transaction Costs –
Positive Up / Downstream Capability Correlation, with Expansion Costs (in both Segments)
Figure 4a: Aggregate Specialization & Transaction Costs –
Negative Up / Downstream Capability Correlation, with Expansion Costs (only Downstream)

Figure 4b: Aggregate Specialization & Transaction Costs –
Zero Up / Downstream Capability Correlation, with Expansion Costs (only Downstream)

Figure 4c: Aggregate Specialization & Transaction Costs –
Positive Up / Downstream Capability Correlation, with Expansion Costs (only Downstream)