Delay Cost, Knowledge Hierarchy, and Wages*

CHENG CHEN
University of Hong Kong

WING SUEN
University of Hong Kong

April 27, 2016

Abstract. We provide a new approach to study optimal design of knowledge hierarchies under general assumptions. When problem-solving becomes more urgent, employees at each level solve more complex problems and earn more. The organization structure becomes flatter, with fewer layers but a larger span of control at each layer. Moreover, knowledge acquisition is disproportionately concentrated among lower level employees, which results in shrinking intra-firm wage differentials across layers. We find that labor productivity of the firm increases after delay becomes more costly, despite the direct output loss due to greater delay. Using Colombia plant-level data, we find that exporting firms—which face higher delay cost—have fewer layers, larger span of control, and pay higher wages at all levels.

Keywords. knowledge complementarity; delay and exporting; organization design; firm productivity; export-wage premium

JEL Classification. D21; D23; F12; L22; L23

*We are grateful to Pol Antràs, David Blau, Davin Chor, Robert Gibbons, Jin Li, Kevin Lang, Dilip Mookherjee, Arijit Mukherjee, Andrew Newman, Chang Sun, Stephen Terry, Zhigang Tao, HuanXing Yang, and seminar participants at BU, MIT, MSU, OSU, PKU and USC for their valuable comments. Financial support from HKU is greatly appreciated. Finally, Cheng Chen thanks IED of Boston University for its hospitality when the paper was written.

1. Introduction

Business decisions are often time-critical. That is why floor traders are given a lot of leeway to make decisions on the spot. The importance of delay cost for firm outcomes and organization design is not confined to the finance sector. Galbraith (1977) found that "after 1964 the problem facing Boeing was not to establish a market but to meet the opportunities remaining as quickly as possible. Now a delay of a few months would result in canceled orders and fewer sales." He reported that "to respond to competitive time pressure from Douglas, Lockheed, and the British-French Concorde, Boeing was forced to drastically reduce the time devoted to product development and design." Whitney (1988) showed how fast an organization completes its task substantially affects its profit and revenue. Furthermore, Rajan and Wulf (2006) explained that one major reason for increased decentralization inside U.S. firms in recent years was to "enabl[e] faster decision making and execution." In addition, Bloom, Van Reenen and Sadun (2010) argued that "tougher competition may make local manager's information more valuable, as delays to decisions become more costly." Using firm-level survey data, they found that in more competitive markets, which are associated with high delay cost, firms are more decentralized. In short, existing evidence shows that increased delay cost plays a key role in shaping organizational structures of the firm.

Theoretical literature on how the delay cost affects firm structure and efficiency is scant. There is a literature starting from Radner (1992, 1993) that explores the optimal organizational structure that minimizes the processing time for a given task.[1] However, some of Radner's results would imply rather strange organizational structures that are hardly consistent with reality. Moreover, important questions such as how increased demand for fast decision making affects firm organization and productivity, knowledge and wages of workers are left unanswered. In this paper, we provide a model that incorporates the cost of delay explicitly to answer these questions.

We build on the framework of Garicano (2000) and Garicano and Rossi-Hansberg (2004, 2006, 2012), which model the firm as a knowledge-based hierarchy. In order to realize potential output, agents have to solve problems arising from the production process, which require them to acquire knowledge. Hierarchies are created to economize on the cost of acquiring knowledge: subordinates learn less knowledge and deal with routine problems, while referring the complex but infrequent problems to their more knowledgeable supervisors. We depart from this model by introducing delay cost. Problems have to be solved, and solved quickly. Different from Garicano (2000), we assume that if a problem is solved by agents higher up in the hierarchy, it causes a greater loss in revenue due to delay in time. Unsolved problems at the top layer generate losses in revenue as well. The rationale for our specification is that output or sales are affected by how fast an organization carries out the production. Naturally, the slower an organization makes decisions or solves problems, the more losses

[1] Other papers include Bolton and Dewatripont (1994) and Patacconi (2009).

it bears.[2]

The key insight from our framework is that increased delay cost makes the firm less hierarchical, more knowledge-intensive, and more productive. In order to prevent costly delay, the firm recruits more knowledgeable workers at all levels so that there are fewer problems waiting to be solved by upper management. More importantly, since delay is cumulative, making agents at lower levels acquire knowledge is relatively more effective in reducing delay than making middle or upper level workers acquire knowledge. As a result, the firm invests disproportionately on knowledge acquisition by lower level workers—they are more empowered in the production process. Since more knowledgeable workers command higher wages, intra-firm wage inequality measured by the wage ratio between two adjacent layers shrinks. This prediction is starkly different from the effect of improved information and communication technology on wages.[3]

An increase in delay cost tends to reduce firm size, because it is a negative shock for the firm. However, it tends to increase firm labor productivity (revenue per employee). The key to understanding this is that employees at each layer acquire more knowledge as delay becomes more costly, which increases output conditional on the number of employees. Furthermore, this effect is magnified in a knowledge hierarchy through complementarity: as lower level workers can deal with more complex problems themselves, their supervisors can manage a greater span of control. In other words, when delay becomes more costly, the firm uses fewer yet more knowledgable agents to produce. This result helps explain why small and medium-sized exporting firms (which faces more volatile demand conditions and hence higher delay costs) may actually be more productive than big non-exporters.

Using Colombia plant-level data, we provide evidence in Section 2 to support our model's predictions on internal firm organization and wages. We find that, conditional on other firm-level characteristics, exporting firms, which are likely to lose more due to delay, have fewer layers, larger span of control, and pay higher wages at all layers. In reality, exporting is an activity whose success is particularly sensitive to the delay of transactions. Cavusgil et al. (2012) argued that "exporting requires the ability to adapt rapidly to foreign market opportunities and faster decision making and implementation of new operating methods." They documented that the emergence of less hierarchical small and medium-sized exporters in recent years was a response

[2]One possible economic story for our setup is the following. First, demand is assumed to change every period after the firm begins to deal with a problem. If the problem is solved by bottom layer workers at the current period, then what is produced is sold without loss. Second, we assume that it takes i periods to get the problem fixed by workers at layer i, and the firm loses a fixed fraction of output when the demand changes. Then, the probability that the demand has vanished increases with time. As a result, the firm loses more output when the problem is solved at upper layers. Finally, unsolved problems also generate losses due to both delay and the damage on the firm's reputation.

[3]Improved information technology makes all agents learn more without generating a relative wage effect. Improved communication technology empowers agents at upper layers as they become more effective in communicating with subordinates.

to the concern about undue delay in exporting. Complementing their anecdotal evidence, we provide statistical evidence to show that exporting firms indeed have less hierarchical structures and pay higher wages. These findings are robust to different definitions of what constitute a "layer" within the firm, and to the use of instrumental variables for exporting status. These new findings suggest that the pure effect of exporting on firm organization should be distinguished from the size effect and the productivity effect.

Our theory helps explain the difference in management practices between Japanese firms and U.S. firms. As Aoki (1989, 1990) documented, workers on the production floor in Japan have higher authority and deal with more complex problems compared with their U.S. counterparts. At the same time, one key element of Toyota's lean production is to fasten the firm's decision making process and make the firm response to changes in market environment more rapidly. Our theory shows that if a firm cares more about the delay in its production and decision-making process, it should make workers at lower hierarchical levels acquire more knowledge and empower them more (i.e., compared with middle- and high-level managers). Therefore, our delay story is useful for us to better understand these observations.

This paper contributes to the literature on internal firm organization in two ways. First, it proposes a new approach to solving for optimal organizational choices of the firm. Our approach exploits two key insights: (1) knowledge at different levels in a hierarchy are complementary; and (2) a hierarchical structure entails (conditional) separability that simplifies the analysis of linkages across different layers. We elaborate on these key points in Section 3. Using results from monotone comparative statics (Topkis 1998), our approach is more transparent and imposes fewer restrictions on functional forms, while existing work (Caliendo and Rossi-Hansberg 2012; Chen 2015) relies on distributional assumptions to solve for optimal organizational choices. We also obtain new results regarding optimal hierarchical structures that are previously available only through numerical simulations. Our approach can be applied to study other interesting questions, such as how improved information technology and the productivity draw of the firm affect the optimal number of layers. Second, our paper uncovers distinctive effects of increased delay cost on intra-firm wage inequality and firm productivity. These predictions can be readily confronted with the data; we leave them for future research.[4]

Our work also contributes to the literature on trade and internal firm organization. The existing literature documents that exporting firms have more layers than non-

[4]Beggs (2001) also discussed how increased delay cost affects firm hierarchy, and our paper differs from it in two fundamental ways. First, we use the supermodularity approach to systematically solve for the optimal organizational structure under few assumptions. Furthermore, our new approach uncovers two key properties (of the profit function) that are common across several hierarchy models, which Beggs (2001) is silent on. Second, a key contribution of our paper is to explore how delay cost affects intra-firm wage inequality and firm productivity, while Beggs (2001) focused exclusively on how delay cost affects internal firm structure and absolute wages.

exporting firms (e.g., Tag 2013; Caliendo, Monte and Rossi-Hansberg 2015; Frederic 2015). This may be due to the size-effect or the productivity effect, since exporting firms tend to be bigger and more productive than non-exporters. The result we uncover is that, conditional on firm size and productivity, exporting firms actually have fewer layers on average. This new finding complements existing empirical studies. Furthermore, it points out one additional channel (i.e., the motive to reduce losses due to delay) through which exporting affects internal firm organization, which deserves more attention for future research.

2. Empirical Findings

Our model offers distinctive predictions concerning the effects of delay cost on the firm organization and wages. Using exporting status as a proxy for firms that face higher delay cost, our model predicts that exporting firms (1) have fewer number of layers in the firm hierarchy; (2) have larger span of control at each layer; and (3) pay higher wages. We use Colombia plant-level data to illustrate these predictions and admit that the empirical findings we are going to present are *not* causal evidence for our theory.

The Colombia plant-level data contain survey information between 1981 and 1991 on all registered manufacturing plants in Colombia with more than ten employees.[5] It has been extensively used in previous research (e.g., Roberts 1996; Roberts and Tybout 1997; Fernandes 2007).[6]

The data set contains employment and wage information for agents at various levels of the firm hierarchy: (a) owners; (b) managers; (c) technicians and skilled workers; and (d) unskilled workers, including apprentices. Our basic measure of the number of *layers* excludes the owners and counts whether each of the remaining three types (b), (c), and (d) are present in the firm. The maximum value of *layers* is 3 and the minimum is 0, with a sample mean of 2.51. Our basic measure of the *span* of control of managers is the ratio of the number of agents in category (c) to those in categories (a) and (b).[7] The mean value of *span* is 8.3, which is close to the value found by Guadalupe and Wulf (2010).

The average number of agents in the three categories we identify are 2, 18, and 48. Our construction of firm hierarchy can be mapped into the structure of top managers, middle-level managers, and workers used by most companies in the world. Average wage and employment (by layer) defined in this paper do follow a hierarchical pattern: the number of workers decreases while average wage increases when we move up along the ladder of layers. The extent of variation in the *layers* variable across plants

[5] In several years, very few plants with employment less than ten are also included.

[6] Special thanks are given to Prof. Stephen Redding for sharing the data with us. For summary statistics of the data set, see the Appendix A.1.

[7] A substantial fraction of Colombian plants had owners who run their firms. As a result, many plants have zero managers in the data.

and time is not small, we will discuss this in what follows.

2.1. Exporting Status and the Number of Layers

The basic equation is an ordinary least squares regression between $layers_{it}$ and $export_{it}$ (dummy variable equal to 1 if plant i engages in exporting in year t). We control for the following firm-level characteristics: logarithm of employment (*employment*), share of skilled workers in total employment (*skill-share*), logarithm of deflated value-added per worker (*productivity*), and firm age (*age*). We use one-period lag of the above four variables to avoid the most obvious form of reverse causality. We use firm fixed effects to capture all time-invariant firm-level unobservables that affect the choice of the number of layers. Industry-year fixed effects and location fixed effects are included into the regression, as previous research has identified that import tariffs and competition at the industry-year level affect firm's decentralization decisions (Acemoglu et al. 2007; Guadalupe and Wulf 2010). A dummy variable $import_{it}$ for importing status is also included, since advanced technologies acquired through importing may make these firms adopt different internal organization.

The first column of Table 1 shows the ordinary least squares regression results. Firm size, the share of skilled workers, and labor productivity positively affect the number of layers. Age indicator is positive, although not statistically significant. These findings are consistent with the existing literature (Caroli and Van Reenen 2001; Bresnahan et al. 2002; Acemoglu et al. 2007; Tag 2013; and Caliendo, Monte and Rossi-Hansberg 2015). What is interesting is that exporting status is associated with fewer layers after we control for these firm-level variables. Importing status, on the other hand, does not have a significant effect on *layers*.[8]

Having implemented our regression analysis, we explore how Colombian plants had changed their organizational structure after starting or stopping to export. We report three key findings here; interested readers are referred to Appendix A.2 for details. First, a substantial number of plants—about 10 percent of observations (i.e., plant-year pairs)—had changed the number of layers in two consecutive years. This is surprising, since reorganization through adding or dropping layers is usually thought to take a long time. Second, when Colombian plants changed the number of layers, they usually added or deleted the layer of managers or technicians. This is what we expect—firms add and drop layers mainly at the level of managerial staffs rather than skilled or unskilled workers. Finally, when we compare plants that started to export with plants that did not change their exporting status, exporting starters reduced their number of layers although their firm size increases on average (in the relative sense). This finding explains why exporting status has a negatively significant impact on the number of layers, since our identification for this regressor comes from plants that had

[8]The classical measurement error for the number of layers would not make our estimates be biased, since the number of layers is the *dependent variable* of our estimation equation.

Table 1. *Effects of Exporting Status on Organizational Structure and on Wages at Various Levels of the Hierarchy*

	(1) *layers*	(2) *span*	(3) *wage* (unskilled)	(4) *wage* (skilled)	(5) *wage* (technicians)	(6) *wage* (managers)
$export_{it}$	-0.0233^{***}	0.141	65.01^{***}	84.82^{***}	127.70^{***}	440.50^{***}
	(-3.04)	(0.61)	(5.31)	(4.44)	(2.80)	(4.14)
$import_{it}$	0.0081	-0.032	-4.80	-14.95	23.12	126.70
	(0.94)	(1.37)	(-0.51)	(-0.99)	(0.60)	(1.47)
$employment_{it-1}$	0.1180^{***}	1.710^{***}	-8.83	37.91^{***}	69.61	118.00
	(9.39)	(11.04)	(-0.97)	(3.09)	(1.54)	(1.56)
$skill\text{-}share_{it-1}$	0.0441^{*}	-	29.72	-30.50	-482.60^{***}	30.06
	(1.83)		(1.57)	(-0.84)	(-3.62)	(0.13)
$productivity_{it-1}$	0.0246^{***}	0.425^{***}	38.70^{***}	73.73^{***}	124.50^{***}	231.80^{***}
	(3.25)	(5.87)	(7.22)	(6.06)	(4.55)	(3.88)
age_{it-1}	0.0044	0.016	-44.38^{***}	-79.44^{***}	-202.40^{***}	-551.60^{***}
	(0.94)	(0.95)	(-6.14)	(-6.33)	(-5.50)	(-7.56)
no. obs.	55031	54600	54001	50671	20051	35915
adjusted R^2	0.735	0.743	0.815	0.759	0.731	0.667

The *age* indicator equals 0 if firm age is less than or equal to 5; equals 1 if between 6 and 10; equals 2 if between 11 and 20; equals 3 if between 21 and 40; and equals 4 if above 40. Standard errors are clustered at the 5-digit SIC level. All regressions include firm fixed effects, industry-year fixed effects, and location fixed effects. t-statistics are in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

changed exporting status across years.[9]

We have implemented two robustness checks and found similar results as in Table 1. First, we adopt an alternative measure of number of layers (*layers-2*) by separating technicians from skilled workers and use it as the dependent variable. The result is qualitatively the same. Second, we use the share of exports in total sales as a proxy for exporting status and rerun all the regressions. Again, the results are qualitatively the same. As we mention in the introduction, timeliness of decisions is likely to be more important for export sales than for domestic sales. We conjecture that exporting firms have fewer layers in the hierarchy in order to reduce losses due to delay.[10]

2.2. Exporting Status and the Span of Control

In this subsection, we show that exporting status has a positive impact on managerial span of control. The dependent variable is *span*, the number of technicians and skilled

[9]We also find that, compared with plants that did not change their exporting status, exporting stoppers also reduced the number of layers. This is mainly due to the substantial decrease in firm size when plants stopped exporting.

[10]Colombia was very poor in the 1980s, with a per capita real GDP of 1160 USD in 1985. Communications technology was unlikely to have been advanced substantially during the sample period. If better communications technology were the true reason for exporting firms' delayering, we would expect that other things being equal, production workers earn less in those firms—a key implication of the knowledge hierarchy model. Our following empirical analysis shows that the opposite pattern.

workers divided by the number of managers. Since the number of skilled workers enters directly into the calculation of span, we exclude *skill-share* from the independent variables to avoid superfluous correlation between dependent and independent variables. Other specifications are the same as the regressions for *layers* in the previous subsection.

Results for the OLS regression are presented in the second column of Table 1. In this regression, *export* has a positive impact on *span*, although it is not statistically significant. We have also used *span-2* as the dependent variable, defined as the ratio of skilled workers to managers and owners. The estimated coefficient on *export* is slightly larger, though it is still statistically insignificant.

2.3. Exporting Status and Wages

In this subsection, we show that Colombian plants pay higher wages to workers at all hierarchical layers after they become exporting firms. We use the average wage of employees at various hierarchical levels as the dependent variables. Regression results are presented in columns (3)–(6) of Table 1. It is evident from these tables that, workers at all hierarchical layers receive higher wages after their firms start to export. Moreover, compared with the exporting status, the importing status has a much smaller and sometimes opposite effect on wages. We conjecture that a technology story does not explain these wage patterns very well, since both exporting and importing firms probably have more advanced technologies.

Based on the above three empirical findings, we conjecture that it is the difference in the cost of delay that contributes to the observed difference between exporters and non-exporters. Cavusgil et al. (2012) argued that rapid adaptation to foreign markets and fast decision making are crucial to the success of international business. Shipping goods to foreign markets takes much longer time than shipping them domestically. This is not only due to the distance, but also due to the long time spent on getting goods through customs (Sabur et al. 2004; Djankov, Freund and Pham 2010). Since demand changes rapidly, it generates more losses to exporting firms if decision making is delayed. In what follows, we build a theoretical model to show how the cost of delay affects internal firm organization and wages.

3. A Model of Delay Cost in Knowledge Hierarchies

We adapt Garicano's (2000) model of knowledge hierarchies to study the effects of delay cost on organizational design. The production process presents problems which have to be solved by the firm. The difficulty of a problem is indexed by Z, which is a random draw from distribution F on the support $[0, \infty)$. The distribution is continuous with density f.[11] Workers are organized according to a hierarchy: if a problem

[11] To highlight the insight that lower-level workers deal with more common problems, Garicano (2000) orders problems in such a way that the frequency of their appearance is decreasing. Our model

cannot be solved by a worker at a certain level, he refers it to his supervisor immediately above him. Knowledge in problem solving is assumed to be cumulative, in the sense that supervisors have to know what their subordinates know. Let the knowledge of a worker at level i be represented by z_i. Because of the hierarchical structure of problem-solving, production workers (level-0 workers) solve problems in the range $[0, z_0]$, level-1 supervisors solve problems in the range $(z_0, z_1]$, and so on. In general, a problem with $Z \in (z_i, z_{i+1}]$ has to go through i layers (including the bottom layer) before it is finally solved by a level-i supervisor. If the highest level supervisor in the firm is L, then a problem with $Z > z_L$ will remain unsolved after passing through the entire hierarchy.

Each production worker is paired up with one unit of capital to deal with one problem. Each supervisor can deal with $1/h$ problems referred to him by his subordinate (and does not require any capital). We assume that $h \leq 1$. Let n_i represent the number (or mass) of workers at level i. With probability $1 - F(z_0)$, a production worker cannot handle a problem and has to refer it to his supervisor. Thus, the total number of level-1 supervisors required is $n_1 = n_0 h(1 - F(z_0))$. As a result, the *span of control* of a level-1 supervisor is:

$$s_1 = \frac{n_0}{n_1} = \frac{1}{h(1 - F(z_0))}.$$

The assumption that $h \leq 1$ ensures that $s_1 > 1$, which is consistent with observed data in hierarchical firms (e.g., Caliendo, Monte and Rossi-Hansberg (2015)).

More generally, the probability that a level-i has to deal with a problem referred to him is $1 - F(z_{i-1})$. Thus, the total number of level-i supervisors required is $n_0 h(1 - F(z_{i-1}))$. For $i \geq 2$, the span of control is given by:

$$s_i = \frac{n_{i-1}}{n_i} = \frac{1 - F(z_{i-2})}{1 - F(z_{i-1})}.$$

Note that the parameter h does not affect s_i for $i \geq 2$ directly.

The value of production if a problem is successfully solved is A. Thus, in a firm with n_0 production workers and L layers of supervisory workers, the potential total output is $n_0 A F(z_L)$. Total wage cost is $\sum_{i=0}^{L} n_i w_i$. More knowledgeable workers are more expensive to hire. We assume that $w_i = \omega + c z_i$, with $\omega \geq 0$ and $c > 0$. Because of the fixed proportions technology, the firm requires n_0 units of capital. To obtain a determinate firm size, we assume that the total cost of capital is $C(n_0)$, where $C(\cdot)$ is strictly convex. The key departure from the standard knowledge hierarchy model is that we introduce a cost due to delay, represented by $n_0 D_L$ (where D_L is the delay cost per production worker for a firm with L layers) and to be elaborated below. Total

allows for the possibility that more difficult problems are not necessarily less common. We do not require the assumption that the density function f is non-increasing.

profit of the firm is equal to:

$$\Pi = n_0 AF(z_L) - n_0 D_L - \sum_{i=0}^{L} n_i w_i - C(n_0). \tag{1}$$

As a convention, we say that a firm has L layers if there are L levels of supervisors plus one level of production workers. The problem for the firm is to choose the number of layers L and the number of workers n_i ($i = 0, \ldots, L$) at each layer, together with their knowledge level z_i, to maximize (1) subject to the labor requirement constraints $n_i = n_0 h(1 - F(z_{i-1}))$ for $i = 1, \ldots, L$.

Problems that are solved quickly by lower level employees are worth more than those that are solved slowly by higher level employees after going through the firm hierarchy. Specifically, we assume that the firm incurs an additional delay cost of $A\varphi$ every time a problem has to go up one level in the knowledge hierarchy, where φ is the *marginal* cost of delay.[12] Thus, problems are solved by level-i workers are yield a realized output of $A(1 - i\varphi)$. Problems that cannot be solved even by level-L yields no output, but the firm still has to bear a delay cost of $AL\varphi$. With these assumptions, delay cost per production worker is given by:

$$D_L = A\varphi \left(L(1 - F(z_L)) + \sum_{i=1}^{L} i(F(z_i) - F(z_{i-1})) \right) = A\varphi \left(\sum_{i=1}^{L} 1 - F(z_{i-1}) \right). \tag{2}$$

3.1. Supermodularity and Conditional Separability

Substituting the labor requirements n_i ($i = 1, \ldots, L$) and the delay cost D into equation (1), the firm's maximization problem can be written as:

$$\max_{n_0, L, \{z_0, \ldots, z_L\}} \Pi = n_0 \pi_L(z_0, \ldots, z_L) - C(n_0), \tag{3}$$

where

$$\pi_L(z_0, \ldots, z_L) = AF(z_L) - (\omega + cz_0) - \sum_{i=1}^{L} (1 - F(z_{i-1}))(A\varphi + h(\omega + cz_i)) \tag{4}$$

is profit per production worker.

Since $\pi_L(\cdot)$ is independent of n_0, the firm's maximization problem can be analyzed in three steps. First, holding the number of layers fixed at L, the firm chooses the optimal knowledge level z_i ($i = 0, \ldots, L$) at each layer to maximize profit per production worker. Second, we study the optimal choice of the number of layers L. Finally, we

[12] It takes much longer time to ship goods abroad than shipping them domestically. Thus, conditional on problems' being solved at the same layer, the probability with which demand changes is higher for exporting firms than for non-exporting firms. This leads to a bigger φ for exporting firms, which drives the empirical patterns in organization structure and wages described in Section 2.

determine the optimal firm size by choosing the number of production workers n_0. Given n_0 and z_i for $i = 0, \ldots, L$, the number of supervisory workers n_i ($i = 1, \ldots, L$) can be obtained through the labor requirement constraints.

Before we proceed with the analysis, we make two observations about the function $\pi_L(\cdot)$ in equation (4) that are crucial for our results. First, $\pi_L(\cdot)$ is *supermodular* in $(z_0, \ldots, z_L)$, which reflects knowledge complementarity within the organization. The marginal benefit of raising z_i stems from the fact that fewer problems need to be referred to supervisors at level $i+1$. The higher is z_{i+1}, the more expensive are these supervisors, and hence the greater is the benefit from increasing z_i. The cost of raising z_i depends on the marginal cost of knowledge acquisition, c, and on the number of level-i workers hired, n_i. The higher is z_{i-1}, the fewer the number of level-i workers needed, and hence the smaller is the marginal cost of raising z_i. Thus, we have $\partial^2 \pi_L / \partial z_i \partial z_j \geq 0$ for all $i \neq j$.

Second, thanks to the hierarchical form of organization, knowledge complementarity is confined to workers between successive layers of the firm. In other words, $\partial^2 \pi_L / \partial z_i \partial z_j = 0$ whenever $|i - j| > 1$. An implication of this observation is that the optimal knowledge composition for workers above some level i is related to the optimal knowledge composition for workers below level i only through z_i. Conditional on z_i, the optimal structure of the hierarchy above level i can be determined separately from that below level i. Specifically, for any $i < L$, we can write

$$\pi_L(z_0, \ldots, z_L) = \pi_i(z_0, \ldots, z_i) + \Delta^{i,L}(z_i, \ldots, z_L), \tag{5}$$

where

$$\Delta^{i,L}(z_i, \ldots, z_L) = AF(z_L) - AF(z_i) - \sum_{j=i+1}^{L} (1 - F(z_{j-1}))(A\varphi + h(\omega + cz_j)). \tag{6}$$

The function in equation (6) represents the difference in profit per worker if the firm adds $L - i$ layers on top of its existing structure of i layers. Note that profit from the bottom part $\pi_i(\cdot)$ and profit from the upper part $\Delta^{i,L}(\cdot)$ are related to each other only through z_i. We will take advantage of such *conditional separability* in much of the subsequent analysis.

The above two key properties are also present in other types of hierarchy models such as monitoring-based hierarchy models (Calvo and Wellisz 1978; Qian 1994; Chen 2015) and hierarchy models based on queueing theory (Beggs 2001). Thanks to this observation, our new approach can be used to study how other factors (e.g., monitoring technology, firm productivity) affect internal firm organization under few functional form assumptions. We do not elaborate on this point in the current paper, although more detailed discussions are available upon request.

3.2. Organizational Choices Given the Number of Layers

We solve the firm's optimization problem for fixed n_0 and L first. The first-order conditions for maximizing profit per production worker (4) are:

$$\begin{aligned} f(z_0)(A\varphi + h\omega + hcz_1) - c &= 0; \\ f(z_i)(A\varphi + h\omega + hcz_{i+1}) - hc(1 - F(z_{i-1})) &= 0, \quad i = 1, \ldots, L-1; \\ f(z_L)A - hc(1 - F(z_{L-1})) &= 0. \end{aligned} \tag{7}$$

The set of problems that a level-i worker can deal with is $(z_{i-1}, z_i]$. We say that a level-i worker *deals with more complex problems* if z_{i-1} and z_i both weakly increase. The following proposition characterizes how the marginal cost of delay affects organizational choices of the firm.

Proposition 1. *Holding the number of layers fixed, when the marginal cost of delay increases, (a) workers at every level deals with more complex problems; (b) workers at every level earn higher wages; (c) the span of control of direct supervisors of production workers increases; and (d) the span of control of every level of supervisors increases if* $1 - F$ *is log-concave.*

Proof. The profit function π_L is supermodular in $(z_0, \ldots, z_L)$ and φ. By Topkis's monotonicity theorem, the optimal z_i ($i = 0, \ldots, L$) increases in φ. Workers at level i deal with more complex problems because both z_{i-1} and z_i increases. Workers earn higher wages because $w_i = \omega + cz_i$. Part (c) follows from the fact that $s_1 = 1/(h(1 - F(z_0)))$. Finally, from the first-order condition (7) for level-i supervisors ($i = 2, \ldots, L$), we can write:

$$s_i = \frac{1}{hc} \frac{f(z_{i-1})}{1 - F(z_{i-1})} (A\varphi + h\omega + hcz_i).$$

Since the hazard rate increases in z_{i-1} when $1 - F$ is log-concave, and since z_{i-1} and z_i both increase in φ, the span of control s_i increases in φ as well. ∎

As delay becomes more costly, the organization wants to solve more problems quickly without going through many layers in the hierarchy. The firm becomes more knowledge-intensive by having all agents learn to deal with more complex problems and earn more. Given n_0, the number of supervisory workers at all levels fall as workers become more knowledgeable. But because n_i falls proportionally more than does n_{i-1}, the span of control s_i increases.

3.3. The Optimal Number of Layers

Let $z_i^*(L)$ ($i = 0, \ldots, L$) represent the optimal knowledge levels that maximize profit per production worker for a firm with L layers, and let $\pi_L^* = \pi_L(z_0^*(L), \ldots, z_L^*(L))$ represent the corresponding profit level. Suppose $L' > L$. To compare the profit from having L layers versus L' layers, we make use the conditional separability relation (5) to write $\Delta^{L,L'}(z_L, \ldots, z_{L'})$ as the incremental profit per production worker if the firm

raises the number of layers from L to L'. From equation (6), the incremental profit can be written in the following form:

$$\Delta^{L,L'} = (1 - F(z_L)) \left[A - A\frac{1 - F(z_{L'})}{1 - F(z_L)} - \sum_{j=L+1}^{L'} \frac{1 - F(z_{j-1})}{1 - F(z_L)} (A\varphi + h\omega + hcz_j) \right].$$

The expression in brackets is strictly decreasing in z_L. Thus, for any $(z_{L+1}, \ldots, z_{L'})$, the function $\Delta^{L,L'}(\cdot, z_{L+1}, \ldots, z_{L'})$ is *single-crossing from above* in z_L. Furthermore, it is straightforward to verify that $\Delta^{L,L'}(\cdot)$ is *supermodular* in $(z_L, \ldots, z_{L'})$.

Lemma 1. *If an organization is indifferent between having L layers or L' layers, then (a) level-i workers ($i = 0, \ldots, L$) in the less hierarchical organization deal with more complex problems than their counterparts in the more hierarchical organization; (b) total delay is shorter in the less hierarchical organization; and (c) the span of control at each level $i = 1, \ldots, L-1$ is wider in the less hierarchical organization if $1 - F$ is log-concave.*

Proof. By definition,

$$\begin{aligned} \pi^*_{L'} &= \pi_L \left(z^*_0(L'), \ldots, z^*_L(L') \right) + \Delta^{L,L'} \left(z^*_L(L'); z^*_{L+1}(L'), \ldots, Z^*_{L'}(L') \right) \\ &\leq \pi^*_L + \Delta^{L,L'} \left(z^*_L(L'); z^*_{L+1}(L'), \ldots, z^*_{L'}(L') \right). \end{aligned}$$

Therefore, $\pi^*_L = \pi^*_{L'}$ implies $\Delta^{L,L'}(z^*_L(L'); z^*_{L+1}(L'), \ldots, z^*_{L'}(L')) \geq 0$. Furthermore,

$$\begin{aligned} \pi^*_L &= \pi_{L'} \left(z^*_0(L), \ldots, z^*_L(L), z^*_{L+1}(L'), \ldots, z^*_{L'}(L') \right) - \Delta^{L,L'} \left(z^*_L(L); z^*_{L+1}(L'), \ldots, z^*_{L'}(L') \right) \\ &\leq \pi^*_{L'} - \Delta^{L,L'} \left(z^*_L(L); z^*_{L+1}(L'), \ldots, z^*_{L'}(L') \right). \end{aligned}$$

Therefore, $\pi^*_L = \pi^*_{L'}$ implies $\Delta^{L,L'}(z^*_L(L); z^*_{L+1}(L'), \ldots, z^*_{L'}(L')) \leq 0$. Since $\Delta^{L,L'}$ is single-crossing from above in its first argument, we must have $z^*_L(L) \geq z^*_L(L')$.

Since $z^*_L(L)$ is optimal for L, the optimal choice of knowledge at lower levels in the less hierarchical organization remains unchanged when z_L is fixed at $z_L = z^*_L(L)$. Thus,

$$\left(z^*_0(L), \ldots, z^*_{L-1}(L) \right) = \underset{z_0, \ldots, z_{L-1}}{\arg\max}\, \pi_L \left(z_0, \ldots, z_{L-1}; z^*_L(L) \right).$$

Furthermore, by the conditional separability property, since $z^*_L(L')$ is optimal for L', the optimal choice of knowledge at lower levels in the more hierarchical organization remains unchanged when z_L is fixed at $z_L = z^*_L(L')$. Thus,

$$\left(z^*_0(L'), \ldots, z^*_{L-1}(L') \right) = \underset{z_0, \ldots, z_{L-1}}{\arg\max}\, \pi_L \left(z_0, \ldots, z_{L-1}; z^*_L(L') \right).$$

Since the function $\pi_L(\cdot)$ is supermodular, by Topkis's theorem, $z^*_L(L) \geq z^*_L(L')$ implies $z^*_i(L) \geq z^*_i(L')$ for $i = 0, \ldots, L-1$. This establishes part (a).

Because $L' > L$ and $z_i^*(L') \leq z_i^*(L)$,

$$D_{L'} = A\varphi \left(\sum_{i=0}^{L'} 1 - F\left(z_i^*(L')\right) \right) > A\varphi \left(\sum_{i=0}^{L} 1 - F\left(z_i^*(L)\right) \right) = D_L.$$

This establishes part (b). Part (c) follows from the same logic as in the proof of part (d) of Proposition 1. ∎

Proposition 2. *The optimal number of layers in an organization L is weakly decreasing in the marginal cost of delay φ.*

Proof. Suppose $\varphi' > \varphi$. Let L be the optimal number of layers when the delay cost is φ, and L' be the optimal number of layers when the delay cost is φ'. We want to show that $L' \leq L$. Suppose to the contrary that $L' > L$. By revealed preference, we must have

$$\pi_{L'}^*(\varphi') \geq \pi_L^*(\varphi'), \tag{8}$$

$$\pi_{L'}^*(\varphi) \leq \pi_L^*(\varphi). \tag{9}$$

Furthermore, by the envelope theorem, $\mathrm{d}\pi_L^*(\varphi)/\mathrm{d}\varphi = -D_L/\varphi$. Part (b) of Lemma 1 shows that $D_{L'} \geq D_L$ for $L' > L$ whenever the firm is indifferent between L' and L. This implies that $\pi_{L'}^*(\cdot)$ crosses $\pi_L^*(\cdot)$ at most once and from above. This single-crossing property, together with inequalities (8) and (9), imply that $\varphi' \leq \varphi$, a contradiction. ∎

In Proposition 1, we have already shown that the span of control increases as delay cost rises, if the firm keeps the number of layers constant. But Proposition 2 (together with part (c) of Lemma 1) shows that, even if the firm reduces the number of layers as delay cost rises, the span of control also goes up. These predictions are consistent with the evidence we presented in Section 2, where we showed that Colombian firms adopt a flatter organizational structure (fewer layers and greater span of control) as they start to engage in exporting.

Proposition 2 shows that an organization becomes less hierarchical as delay cost increases. We can strengthen this result to show that such change is "gradual", in the sense that the number of layers falls by one at a time as φ increases continuously. Lemma 1 implies that workers at each level earn more as the firm reduces the number of layers (because $z_i^*(L) \geq z_i^*(L')$ for $L' > L$). We can also strengthen the result to give an upper bound on their wage increase.

To develop these additional results, consider a firm with L layers and with knowledge levels $(z_0, \ldots, z_L)$. We say that it is *unprofitable to add a layer* at the bottom if

$$\delta^0(z_0) \equiv \max_{\zeta \leq z_0} \pi_{L+1}(\zeta, z_0, \ldots, z_L) - \pi_L(z_0, \ldots, z_L) \leq 0. \tag{10}$$

Note that δ^0 depends only on z_0. This follows from the conditional separability property because $\partial^2 \pi_L / \partial z_i \partial z_j = 0$ whenever $|i-j| > 1$. If the firm prefers L layers to $L+1$ layers, then we must have $\delta^0 \leq 0$, but the reverse implication is not true. If $\delta^0 \leq 0$, it is unprofitable to add a layer holding $(z_0, \ldots, z_L)$ fixed, but we may still have $\pi^*_{L+1} \geq \pi^*_L$ when the knowledge levels at all layers are chosen optimally. Similarly, it is unprofitable to add a layer at the top if

$$\delta^L(z_L) \equiv \max_{\zeta \geq z_L} \pi_{L+1}(z_0, \ldots, z_L, \zeta) - \pi_L(z_0, \ldots, z_L). \tag{11}$$

Again, if a firm prefers L layers to $L+1$ layers, we must have $\delta^L \leq 0$.

The following result says that if it is unprofitable to add a layer at the bottom, then it remains so when production workers become less knowledgeable. Likewise, if it is unprofitable to add a layer at the top, then it remains so when top managers become more knowledgeable.

Lemma 2. *(a) $\delta^0(z_0) \leq 0$ implies $\delta^0(z_0') < 0$ for $z_0' < z_0$; and (b) $\delta^L(z_L) \leq 0$ implies $\delta^L(z_L') < 0$ for $z_L' > z_L$.*

Proof. From equation (4) for profit per production worker and from definition (10),

$$\delta^0(z_0) = \max_{\zeta \leq z_0} cz_0 - c\zeta - (1 - F(\zeta))(A\varphi + h\omega + hcz_0),$$

which is strictly increasing in z_0. From equation (5) and from definition (11), $\delta^L(z_L) = \max_{\zeta \geq z_L} \Delta^{L,L+1}(z_L, \zeta)$. We have already shown that $\Delta^{L,L+1}(\cdot; \zeta)$ is single crossing from above, so part (b) of the lemma follows. ∎

Proposition 3. *Suppose both L and $L+1$ layers are optimal at some delay cost φ. Then, $z_i^*(L) \in [z_i^*(L+1), z_{i+1}^*(L+1)]$ for $i = 0, \ldots, L$.*

Proof. We first prove that $z_L^*(L) \leq z_{L+1}^*(L+1)$. Suppose the opposite is true, i.e., $z_L^*(L) > z_{L+1}^*(L+1)$. Then, either (a) there exists some $i < L$ such that $z_i^*(L) \leq z_{i+1}^*(L+1)$; or (b) $z_i^*(L) > z_{i+1}^*(L+1)$ for all $i < L$. In case (a), the conditional separability property implies that

$$\begin{aligned}
(z_{i+1}^*(L), \ldots, z_L^*(L)) &= \arg\max_{\zeta_1, \ldots, \zeta_{L-i}} \Delta^{i,L}(z_i^*(L); \zeta_1, \ldots, \zeta_{L-i}); \\
(z_{i+2}^*(L+1), \ldots, z_{L+1}^*(L+1)) &= \arg\max_{\zeta_1, \ldots, \zeta_{L-i}} \Delta^{i+1,L+1}(z_{i+1}^*(L+1); \zeta_1, \ldots, \zeta_{L-i}).
\end{aligned}$$

Note that $\Delta^{L,L'}(\cdot)$ depends only on the difference $L' - L$ and does not depend on L and L' separately. Furthermore, $\Delta^{L,L'}(\cdot)$ is supermodular. Thus, by Topkis's theorem, $z_i^*(L) \leq z_{i+1}^*(L+1)$ implies $z_L^*(L) \leq z_{L+1}^*(L+1)$, a contradiction. In case (b), we have $z_0^*(L) > z_1^*(L+1) > z_0^*(L+1)$. Since L is optimal, it is unprofitable to add a layer

below $z_0^*(L)$. By Lemma 2, it is unprofitable to add a layer below $z_1^*(L+1)$, which contradicts the optimality of $z_0^*(L+1)$.

Next, we prove that $z_i^*(L) \leq z_{i+1}^*(L+1)$ for all $i < L$. Suppose otherwise. Then, we must have $z_i^*(L) > z_{i+1}^*(L+1)$ for some $i < L$. But then the supermodularity of $\Delta^{i,L}(\cdot)$ implies that $z_L^*(L) > z_{L+1}^*(L+1)$, a contradiction.

We have shown that $z_i^*(L) \leq z_{i+1}^*(L+1)$ for all $i \leq L$. Part (a) of Lemma 1 already established that $z_i^*(L) \geq z_i^*(L+1)$ for all $i \leq L$. The proposition then follows. ∎

Proposition 3 implies that the range of problems solved by level i workers when the firm has L layers (i.e., $(z_{i-1}^*(L), z_i^*(L)]$) partially overlaps with the range of problems solved by the same level of workers when the firm has $L+1$ layers (i.e., $(z_{i-1}^*(L+1), z_i^*(L+1)]$), provided that both organization structures are optimal. Proposition 4 below establishes that if both L and L' are optimal, then (generically) L' cannot differ from L by more than one. Recall that the number of layers is weakly decreasing in the delay cost. An implication of Proposition 4 is that the optimal L is a step function falling by one layer at each step as φ increases.

Proposition 4. *Suppose both L and L' are optimal at some φ. Then, generically, $L' = L+1$.*

Proof. Suppose to the contrary that $L' \geq L+2$. Apart from the non-generic case of $L' = L+2$ with $z_i^*(L') = z_{i-1}^*(L)$ for $i = 1, \ldots, L$, there are three possibilities.

Case (1). $z_{L'-1}^*(L') > z_L^*(L)$. Since it is unprofitable to add a layer above level L when $z_L = z_L^*(L)$, it is unprofitable to add a layer above level $L'-1$ at $z_{L'-1} = z_{L'-1}^*(L')$, which contradicts the optimality of $z_{L'}^*(L')$ under organization structure L'.

Case (2). $z_1^*(L') < z_0^*(L)$. Since it is unprofitable to add a layer below level 0 when $z_0 = z_0^*(L)$, it is unprofitable to add a layer below level 1 when $z_1 = z_1^*(L')$, which contradicts the optimality of $z_0^*(L')$ under organization structure L'.

Case (3). Neither (1) nor (2) is true, i.e., $z_0^*(L) \leq z_1^*(L')$ and $z_L^*(L) \geq z_{L'-1}^*(L')$. In this case, there are $L+1$ layers (including layer 0 and layer L) between $z_0^*(L)$ and $z_L^*(L)$ under organization structure L, and there are $L'-1 \geq L+1$ layers (including layer 1 and layer $L'-1$) between $z_{L'-1}^*(L')$ and $z_1^*(L')$ under organization structure L'. Because there are more layers between a narrower range of knowledge levels in organization L', there must exists $j \in \{1, L\}$ such that layers j and $j+1$ in organization L' are contained in $[z_{j-1}^*(L), z_j^*(L)]$. Define

$$Q(\bar{z}) = \max_{q_0,\ldots,q_{j-1}} \{\pi_j(q_0, \ldots, q_{j-1}, \bar{z})\} - \max_{q_0,\ldots,q_j} \{\pi_{j+1}(q_0, \ldots, q_j, \bar{z})\}$$

to be the difference in profits if the organization has j layers rather than $j+1$ layers, conditional on the top layer having knowledge level $\bar{z}$. Let $q_{j-1}^*(j; \bar{z})$ be the optimal knowledge of the level below $\bar{z}$ when the organization has j layers, and let $q_j^*(j+1; \bar{z})$ be the optimal knowledge of the level below $\bar{z}$ when the organization has $j+1$ layers.

Supermodularity of $\pi_j(\cdot)$ implies that, for all $\bar{z} < z_j^*(L)$,

$$q_{j-1}^*(j;\bar{z}) < q_{j-1}^*(j;z_j^*(L)) = z_{j-1}^*(L).$$

Similarly, supermodularity of $\pi_{j+1}(\cdot)$ implies that, for all $\bar{z} > z_{j+1}^*(L')$,

$$q_j^*(j+1;\bar{z}) > q_j^*(j+1;z_{j+1}^*(L')) = z_j^*(L').$$

Combining these two inequalities gives $q_{j-1}^*(j;\bar{z}) < q_j^*(j+1;\bar{z})$ for all $\bar{z} \in [z_{j+1}^*(L'), z_j^*(L)]$. By the envelope theorem,

$$Q'(\bar{z}) = hc\left(F(q_{j-1}^*(j;\bar{z})) - F(q_j^*(j+1;\bar{z}))\right).$$

Therefore, $Q(\bar{z})$ is strictly decreasing for $\bar{z} \in [z_{j+1}^*(L'), z_j^*(L)]$. Now, since organization structure L is optimal, profits cannot be increased by adding a layer below $z_j^*(L)$ (while keeping everything above layer j fixed). This implies that $Q(z_j^*(L)) \geq 0$. Since $Q(\cdot)$ is strictly decreasing in the region $[z_{j+1}^*(L'), z_j^*(L)]$, this implies $Q(z_{j+1}^*(L')) > 0$. But then this implies that organization structure L' is not optimal (profits could be increased by removing one layer below $j+1$ while keeping everything above $j+1$ fixed), a contradiction.

The reason why we call the case in which $L' = L+2$ and $z_i^*(L') = z_{i-1}^*(L)$ for $i = 1, \ldots, L$ a non-generic case is that there are $L'+1$ unknown variables and $L'+3$ equilibrium conditions for this case. On the one hand, the fact that it is optimal for the L'-layer organization to add the zeroth layer with $z_0^*(L')$ below the first layer with $z_1^*(L')$ implies that

$$cz_1^*(L') - cz_0^*(L') - (1 - F(z_0^*(L'))(A\varphi + h\omega + hcz_1^*(L')) \geq 0$$

On the other hand, the fact that it is *not* optimal for the L-layer firm to add a layer with $z_0^*(L')$ below the current zeroth layer with $z_0^*(L)$ implies that

$$\begin{aligned} & cz_0^*(L) - cz_0^*(L') - (1 - F(z_0^*(L'))(A\varphi + h\omega + hcz_0^*(L)) \\ = \; & cz_1^*(L') - cz_0^*(L') - (1 - F(z_0^*(L'))(A\varphi + h\omega + hcz_1^*(L')) \leq 0, \end{aligned}$$

since $z_1^*(L') = z_0^*(L)$. Thus, we end up with

$$cz_1^*(L') - cz_0^*(L') - (1 - F(z_0^*(L'))(A\varphi + h\omega + hcz_1^*(L')) = 0.$$

Using the same argument, we can derive a condition relating $z_{L'}^*(L')$ to $z_{L'-1}^*(L')$ similar to the above equality. In total, other than equation (7), we have two extra equilibrium conditions implied by $L' = L+2$ and $z_i^*(L') = z_{i-1}^*(L)$ for $i = 1, \ldots, L$. As there is no solution to the optimization problem with $L'+1$ unknowns and $L'+3$ equations in general, the case in which $L' = L+2$ and $z_i^*(L') = z_{i-1}^*(L)$ for $i = 1, \ldots, L$ is a

non-generic case. ∎

The general results we establish in this section are, to the best of our knowledge, new to the literature. Existing work (Caliendo and Rossi-Hansberg 2012; Chen 2015) uses the cost function approach to deal with the problem of optimal organizational design. Such an approach typically requires restrictions on function forms and parameter values of the distribution F. Our approach relies mainly on our observation that the profit function $\pi_L(\cdot)$ is supermodular and conditionally separability—two properties that readily follow from the hierarchical form of the organization structure. We expect that these properties will also feature prominently in other organizational design problems involving hierarchies. We leave these questions for future research.

3.4. Optimal Firm Size

As we remark earlier, the optimal organization design (the number of layers and the knowledge level at each layer) is invariant to the number of production workers n_0. Recall that the production technology requires one unit of capital to be paired with one unit of production worker, and that the cost of capital is given by a convex function $C(n_0)$. To close the model and determine an endogenous firm size, let

$$\pi^*(\varphi) = \max_L \pi_L^*(\varphi)$$

represent the maximum profit per worker when both L and $(z_0, \ldots, z_L)$ are chosen optimally. Since $\pi_L^*(\varphi)$ decreases in φ for any L, $\pi^*(\varphi)$ also decreases in φ. The optimal number of production workers is determined by the following problem:

$$\max_{n_0} \Pi = n_0 \pi^*(\varphi) - C(n_0).$$

Proposition 5. *As the marginal cost of delay increases, (a) total profit of the firm decreases; and (b) the size of the firm in terms of total employment becomes smaller.*

Proof. Total profit decreases because $\pi^*(\varphi)$ is decreasing in φ. The function Π is supermodular in n_0 and $-\varphi$. Thus, the optimal n_0 decreases in φ. Furthermore, by the labor requirement constraint, $n_i = n_0 h(1 - F(z_{i-1}))$. For $i = 1, \ldots, L$, n_i decreases in φ because n_0 decreases and z_{i-1} increases. Thus, total employment falls for any fixed L. At a point when the firm shifts from having $L+1$ layers to having L layers, n_0 does not change (because $\pi_L^*(\varphi) = \pi_{L+1}^*(\varphi)$ at that point) but n_i decreases for all $i = 1, \ldots, L$ (because z_{i-1} increases) and the firm has one fewer layer of supervisory workers. Total employment jumps down discontinuously. ∎

4. Wages and Firm Performance

4.1. Delay Cost, Wages and Intra-firm Wage Inequality

We first summarize what was learned about the level of wages. As delay cost increases, wages at each level i increases both when the firm keeps the number of layers fixed (part (b) of Proposition 1), and when the firm reduces the number of layers (part (a) of Lemma 1). This prediction of our model is consistent with the empirical finding in Section 2, where we presented evidence which shows that workers at every level of the hierarchy receive higher wages after firms start to export.

Now, we explore how increased delay cost affects knowledge acquisition and wages at various layers differently, holding the number of layers fixed. The key finding is a distributional effect: more costly delay causes the firm to disproportionately empower its lower-level employees. The firm becomes more knowledge-intensive at all levels of its hierarchy, but the increase in knowledge is relatively greater among lower levels than among higher levels. Since wages are related to the level of knowledge, this implies that intra-firm wage inequality shrinks as subordinate workers become more empowered relative to supervisors.

The basic intuition for the empowerment effect can be seen from our general model. Take any two levels $i, j \in \{1, \ldots, L-1\}$, and assume $i < j$. The first-order conditions (7) imply:

$$\frac{f(z_i)}{f(z_j)} \frac{A\varphi + h\omega + hcz_{i+1}}{A\varphi + h\omega + hcz_{j+1}} = \frac{1 - F(z_{i-1})}{1 - F(z_{j-1})}.$$

The left-hand-side of the above can be interpreted as the marginal rate of substitution of raising z_i relative to raising z_j. Holding all the knowledge levels constant, the left-hand-side of the above is increasing in φ because $z_{j+1} > z_{i+1}$. In a knowledge hierarchy, the benefit of empowering any given level of workers consists of two parts: (1) a reduction in delay cost; and (2) a saving in wage cost for supervisors. Since low-level supervisors are cheaper than high-level supervisors, the first part (reduction in delay cost) figures more prominently for low level workers than for high level workers, and an increase in delay cost raises the advantage of empowering low-level workers relative to empowering high-level workers.

This intuition is not complete because the marginal rate of substitution depends on hazard rates and other factors as well. To study distributional effects more formally, we need to pay attention to incremental difference in knowledge between successive levels of workers and to impose assumptions about the distribution function F. For $i = 1, \ldots, L$, we define $y_i \equiv z_i - z_{i-1}$. For $i = 0$, define $y_0 \equiv z_0$. Also assume that F is the exponential distribution with parameter λ. Note that the exponential distribution is (weakly) log-concave and entails a constant hazard rate, which greatly simplifies the analysis. With an exponential distribution, the span of control is simply $s_1 = e^{\lambda y_0}/h$, and $s_i = e^{\lambda y_{i-1}}$ for $i = 2, \ldots, L$. Moreover, the first-order conditions (7) can be written

as:

$$\begin{aligned} \lambda(A\varphi + h\omega + hcz_1) &= ce^{\lambda y_0}; \\ \lambda(A\varphi + h\omega + hcz_{i+1}) &= hce^{\lambda y_i}, \quad i = 1, \ldots, L-1; \\ \lambda A &= hce^{\lambda y_L}. \end{aligned} \tag{12}$$

Since z_i is an increasing sequence in i and $h \leq 1$, the first-order conditions (12) imply:

$$y_0 < y_1 < \ldots < y_{L-1}.$$

The following result gives a lower bound to y_1 in terms of primitive parameter.

Lemma 3. *In an optimal organization structure, $y_1 \geq 1/\lambda$.*

Proof. Let L be the optimal number of layers and $(z_0, \ldots, z_L)$ be the corresponding optimal knowledge levels. By definition of optimality, $\pi_L(z_0, z_1, \ldots, z_L) \geq \pi_{L-1}(z_1, \ldots, z_L)$, which is equivalent to $\delta^0(z_1) \geq 0$. Recall that

$$\delta^0(z_1) = \max_{\zeta \leq z_1} cz_1 - c\zeta - e^{-\lambda\zeta}(A\varphi + h\omega + hcz_1).$$

The optimal ζ^* satisfies the first-order condition:

$$e^{-\lambda\zeta^*}(A\varphi + h\omega + hcz_1) = c/\lambda.$$

Thus, $\delta^0(z_1) \geq 0$ implies $cz_1 - c\zeta^* - c/\lambda \geq 0$. But $\zeta^* = z_0$. We therefore have $y_1 = z_1 - \zeta^* \geq 1/\lambda$. ∎

The next result is the key to our analysis of intra-firm wage differentials. It shows that higher delay cost leads to a greater increase in the incremental knowledge of lower-level workers relative to that of higher-level workers.

Lemma 4. *Suppose L is optimally chosen and is fixed as φ varies locally. Then,*

$$\frac{\partial y_0}{\partial \varphi} > \frac{\partial y_1}{\partial \varphi} > \ldots > \frac{\partial y_{L-1}}{\partial \varphi} > \frac{\partial y_L}{\partial \varphi} = 0.$$

Proof. Observe that the each expression on the left-hand-side of the first-order equations (7) is increasing in φ, except for the last equation. We therefore have $\partial y_i/\partial\varphi > 0$ for $i = 0, \ldots, L-1$ and $\partial y_L/\partial\varphi = 0$.

To establish the ranking of the derivatives, we use an induction argument. First, take logarithm on both sides of the first-order conditions (12), and subtract the equation for $i = L-2$ from that for $i = L-1$. We obtain:

$$e^{\lambda y_{L-1}} = e^{\lambda y_{L-2}} + \lambda y_L. \tag{13}$$

Because y_L is fixed respect to φ, and because $y_{L-1} > y_{L-2}$, this equation implies:

$$\frac{\partial y_{L-2}}{\partial \varphi} > \frac{\partial y_{L-1}}{\partial \varphi}.$$

Next, suppose it is true that $\partial y_{k-1}/\partial \varphi > \partial y_k/\partial \varphi$ for some $k \geq 3$. Similar to equation (13), we have the following equation:

$$e^{\lambda y_{k-1}} = e^{\lambda y_{k-2}} + \lambda y_k.$$

Taking logs of both sides and differentiating with respect to φ gives

$$\frac{\partial y_{k-1}}{\partial \varphi} = \frac{e^{\lambda y_{k-2}}}{e^{\lambda y_{k-2}} + \lambda y_k} \frac{\partial y_{k-2}}{\partial \varphi} + \frac{\lambda y_k}{e^{\lambda y_{k-2}} + \lambda y_k} \frac{1}{\lambda y_k} \frac{\partial y_k}{\partial \varphi}, \tag{14}$$

By Lemma 3, $\lambda y_1 \geq 1$, which implies $\lambda y_k > 1$ for all $k \geq 2$. Equation (14) and the induction hypothesis therefore imply $\partial y_{k-2}/\partial \varphi > \partial y_{k-1}/\partial \varphi$.

Finally, we deal with $\partial y_0/\partial \varphi$. The first two equations in (12) can be combined to obtain:

$$he^{\lambda y_1} = e^{\lambda y_0} + h\lambda y_2.$$

We can apply the same reasoning as the above to show that $\partial y_1/\partial \varphi > \partial y_2/\partial \varphi$ implies $\partial y_0/\partial \varphi > \partial y_1/\partial \varphi$. ∎

With the above two lemmas, we can state our main result related to intra-firm wage inequality.

Proposition 6. *Assume $\omega = 0$. When the marginal cost of delay increases and the firm does not adjust the number of layers, the wage ratio w_i/w_{i-1} $(i = 1, \ldots, L)$ between any two successive levels decreases.*

Proof. When $\omega = 0$, $w_i = c\sum_{j=0}^{i} y_j$. Therefore,

$$\frac{w_i}{w_{i-1}} = 1 + \frac{y_i}{\sum_{j=0}^{i-1} y_j}.$$

Since $y_0 < y_1 < \ldots < y_{L-1}$, Lemma 4 implies that the left-hand-side decreases when φ increases. ∎

The assumption that $\omega = 0$ is sufficient but not necessary for Proposition 6 to hold. This assumption is made to ensure that disproportionate knowledge acquisition is reflected in disproportionate wage changes. Furthermore, if we only focus on the top-level managers, we see that

$$\frac{w_L}{w_{L-1}} = 1 + \frac{cy_L}{h\omega + cz_{L-1}}.$$

Since y_L is independent of φ while z_{L-1} is increasing in it, the wage ratio w_L/w_{L-1} necessarily falls even when $\omega > 0$.

4.2. Delay Cost and Firm Performance

Because an increase in φ is a negative shock to the firm, firm profitability unambiguously decreases in φ. The effect of φ on other aspects of firm performance is less obvious. For example, according to equation (2), φ lowers D_L through its effect on z_i (Proposition 1), but it always raises D_L directly. In this subsection, we examine more closely how the cost of delay affects firm performance such as the total loss due to delay, the revenue of the firm, and labor productivity.

Define the revenue per production worker as the output of the firm less the loss due to delay:

$$r_L^*(\varphi) \equiv AF(z_L) - D_L = AF(z_L) - \sum_{i=1}^{L} (1 - F(z_{i-1}))\, A\varphi,$$

where $(z_0, \ldots, z_L)$ are chosen optimally. Total revenue of the firm is simply $n_0 r_L^*$, which is equal to total profit Π less the total wage bill. To study the effect of φ on firm revenue, we make the following assumption.

Assumption 1. *$\lambda(A\varphi + h\omega) \geq 2c$ and $\lambda A \geq 2c$.*

Lemma 5. *Given Assumption 1, the optimal z_1 satisfies (a) $\lambda z_1 > 3$ if $L \geq 2$; or (b) $\lambda z_1 > 2$ if $L = 1$.*

Proof. Consider the expression

$$P(t) \equiv t - \ln\left[\left(\frac{\lambda(A\varphi + h\omega)}{c} + ht\right)\left(\frac{\lambda(A\varphi + h\omega)}{hc} + t\right)\right].$$

The function $P(t)$ is strictly convex, with $P(0) < 0$ and $\lim_{t\to\infty} P(t) > 0$. Therefore, $P(t)$ crosses zero once and from below at some point t^*. By Assumption 1,

$$P(t) \leq t - \ln\left[(2 + ht)\left(\frac{2}{h} + t\right)\right].$$

The right-hand-side of the above is maximized at $h = 2/t$. Substitute this value of h into the equation, we obtain $P(t) \leq t - \ln(8t)$. Since $P(3) \leq 3 - \ln(24) < 0$, we must have $t^* > 3$.

Multiplying the two sides in the first two equations in (12) and taking logarithm, we obtain:

$$\lambda z_1 = \log\left[\left(\frac{\lambda(A\varphi + h\omega)}{c} + h\lambda z_1\right)\left(\frac{\lambda(A\varphi + h\omega)}{hc} + \lambda z_1 + \lambda y_2\right)\right].$$

Since $y_2 > 0$, we must have $P(\lambda z_1) > 0$. Hence, $\lambda z_1 > t^* > 3$. This establishes part (a)

In the case of $L = 1$, we define the function,

$$\hat{P}(t) \equiv t - \ln\left[\left(\frac{\lambda(A\varphi + h\omega)}{c} + ht\right)\left(\frac{\lambda A}{hc}\right)\right].$$

This function crosses zero once and from below at some point $\hat{t}^*$. By Assumption 1, $\hat{P}(2) < 2 - \ln(2 \times 2 + 2 \times 2) < 0$. Since λz_1 satisfies $\hat{P}(\lambda z_1) = 0$, we have $\lambda z_1 = \hat{t}^* > 2$. ∎

Proposition 7. *Suppose Assumption 1 holds. When φ increases and the firm does not adjust the number of layers, (a) revenue per production worker r_L^* decreases; (b) the delay cost per production worker D_L increases; and (c) total revenue $n_o r_L^*$ decreases.*

Proof. We begin with the case $L \geq 2$. Multiply the first-order equation for z_i in (7) by $e^{z_{i-1}}$ and sum over all these equations, we obtain:

$$D_L = \sum_{i=1}^{L} e^{-\lambda z_{i-1}} A\varphi = \frac{c}{\lambda} - e^{-\lambda z_L} A - h \sum_{i=1}^{L} e^{-\lambda z_{i-1}} \left(\omega + c z_i - \frac{c}{\lambda}\right).$$

Therefore, revenue per production worker is

$$r_L^*(\varphi) = A\left(1 - e^{-\lambda z_L}\right) - D_L = A - \frac{c}{\lambda} + h \sum_{i=1}^{L} e^{-\lambda z_{i-1}} \left(\omega + c z_i - \frac{c}{\lambda}\right).$$

A sufficient condition for $r_L^*(\varphi)$ to decrease in φ is that for each $i \geq 1$, $e^{-\lambda z_{i-1}}(\omega + c z_i - c/\lambda)$ decreases in φ. The derivative of the latter expression with respect to φ has the same sign as

$$\left(-\lambda\omega - c(\lambda z_i - 2)\right) \frac{\partial z_{i-1}}{\partial \varphi} + c \frac{\partial y_i}{\partial \varphi}. \tag{15}$$

It follows from Lemma 4 that $\partial y_i / \partial \varphi < \partial z_{i-1} / \partial \varphi$. Thus, the expression in (15) is smaller than

$$\left(-\lambda\omega - c(\lambda z_i - 3)\right) \frac{\partial z_{i-1}}{\partial \varphi}. \tag{16}$$

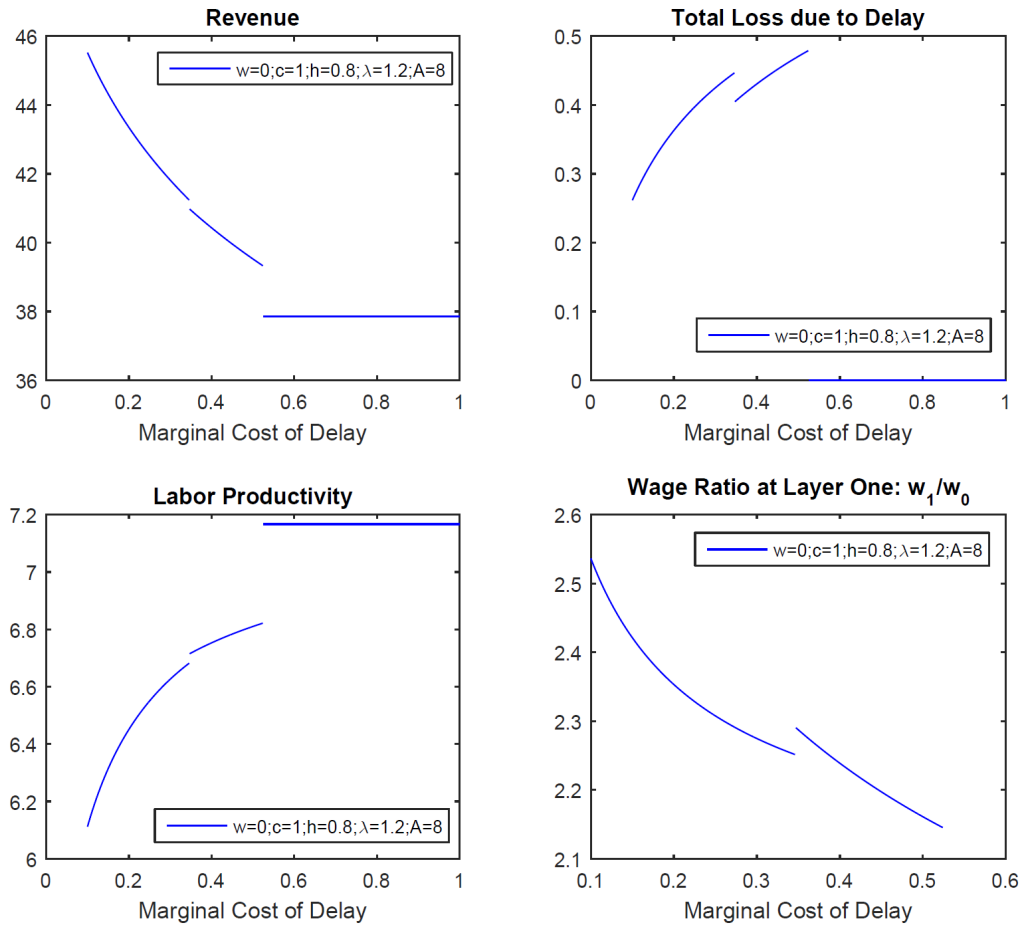
When Assumption 1 holds and when $L \geq 2$, we have $\lambda z_1 > 3$, which implies $\lambda z_i > 3$ for all $i \geq 1$. Thus, the expression (16) is negative, which implies that $r_L^*(\varphi)$ decreases in φ.

In the case $L = 1$, we have $\partial y_1 / \partial \varphi = 0$. Therefore, the expression (15) is negative for $i = 1$ if $\lambda z_1 > 2$, which follows from part (b) of Lemma 5.

To prove part (b), note that $D_L = AF(z_L) - r_L^*$. Since r_L^* decreases in φ while z_L increases in it, we conclude that D_L increases with φ. For part (c), Proposition 5 establishes that n_0 decreases with φ. Therefore, total revenue $n_0 r_L^*$ also falls with φ. ∎

Proposition 7 establishes that higher delay cost lowers the total revenue of the firm.

Figure 1. *Effects of Marginal Cost of Delay on Firm Performance*



But since total employment also decreases (Proposition 5), it is not easy to analytically characterize the effect of delay cost on labor productivity. We resort to numerical simulation to achieve this goal, and the results are presented in Figure 1. In the figure, labor productivity is defined as

$$LP(\varphi, A) \equiv \frac{n_0 r_L^*(\varphi)}{\sum_{i=0}^{L} n_i},$$

where both L and n_i are chosen optimally. For parameter values used in this simulation, only three possible numbers of layers show up in equilibrium: $L = 0, 1, 2$.

As Figure 1 shows, labor productivity increases with the marginal cost of delay. This is because the speed at which total revenue decreases with φ is slower than that for employment. Moreover, when the firm deletes a layer due to an infinitesimally increase in φ, labor productivity jumps up discontinuously. These results are interesting and help us understand why small and medium-sized exporting firms are more productive than big domestic firms.

Figure 1 also reveals that the wage ratio between production workers and their

supervisors jumps up when the firm deletes a layer (due to a small increase in φ). This is because the percentage increase in knowledge stock is larger for employees at upper layers when the firm delayers. This result is interesting, because the wage ratio falls continuously with φ when the firm does not adjust the number of layers. When empirically exploring how wage ratios change after some shock, it is important to take into account whether the firm adjusts the number of layers.

For some parameter values (e.g., for $\varphi < 0.2$) used in the simulations to produce Figure 1 , Assumption 1 does not hold. However, this assumption is sufficient but not necessary for the results stated in Proposition 7. Figure 1 shows that revenue per worker decreases and loss due to delay increases as φ increases even when Assumption 1 fails.

5. Competing Explanations and Further Evidence

The key result of our model is that the marginal cost of delaying negatively affects the number of layers, even after firm productivity and other firm-level characteristics are controlled for. Other arguments for decentralization in firm structure do not seem to fit well into our situation. As argued in Section 2, the story about information and communication technology does not seem to play a big role in the case of Colombia in the 1980s. Second, it might be argued that solving problems arising from exporting activities is more difficult than problems from domestic production (i.e., $c_x > c_d$). If this is the case, exporting firms should make their workers acquire less knowledge, which would imply smaller incremental knowledge stock and smaller span of control of supervisors in exporting firms. However, this result (on the span of control) is inconsistent with what we find in the Colombia data. Finally, it is plausible that problems arising from exporting activities are more unpredictable compared with domestic production. If F is the exponential distribution, this corresponds to saying that $\lambda_x < \lambda_d$. Garicano (2000) has shown that $\lambda_x < \lambda_d$ would imply bottom-layer workers learn less knowledge in exporting firms, since the marginal benefit of acquiring knowledge is smaller for exporting firms. As a result, production workers should receive lower wage, which is inconsistent with our empirical findings. In sum, the difference in the cost of delay (between exporters and non-exporters) generates predictions on the number of layers, the span of control and wages that are more consistent with the data.

We cite existing empirical research to provide further evidence for our theory. First, it is reasonable to assume that the marginal cost of delay is higher in industries with more volatile or uncertain demand. Therefore, according to our theory, firms in such industries should have less hierarchical internal structure, which is exactly what Acemoglu et al. (2007) find. Second, firms in more competitive sectors (e.g., tougher import competition or a lower concentration ratio of sales) probably suffer more from delay. Based on our theory, they should have fewer layers, larger span of control, and more decentralized organization. Bloom, Sadun and Van Reenen (2010) and Guadalupe and Wulf (2010) indeed find these patterns.

Data show that exporting firms are rare, bigger and more productive than non-exporting firms. We can also generate these patterns by incorporating fixed costs and heterogeneity in revenue productivity into the model. Specifically, as in Melitz (2003), if firms differ in revenue productivity, A, and the fixed exporting cost is high enough (relative to the fixed production cost), there is selection into exporting. That is, the most productive firms (among active firms) export, and they are bigger and more productive than non-exporting firms. In short, the stylized facts observed in the trade data can be easily rationalized by our model.

6. Conclusions

Motivated by existing case studies and our own evidence, we set up a knowledge hierarchy model to study how increased delay cost affects optimal firm structure, wages and firm productivity. We provide a new approach to solving for the optimal organizational design under general assumptions, and the model yields several new and testable predictions. First, increased delay cost makes employees acquire more knowledge and earn more. At the same time, the firm increases the span of control and weakly reduces the depth of the hierarchy in order to save the time used to solve problems. Second, increased delay cost generates a distributional effect which disproportionately empowers employees at lower levels to deal with more complex problems. This requires disproportionate increase in investment in knowledge acquisition and results in shrinking intra-firm wage inequality. Third, labor productivity of the firm increases after delay becomes more costly, as firms use more knowledge and less labor to produce. Using Colombia plant-level data and controlling for other firm-level characteristics, we find that exporting firms which are subject to higher delay cost have fewer layers, larger span of control, and pay higher wages at all layers. These findings are consistent with our model's key predictions.

Nevertheless, much remains to be explored. From a theoretical point of view, there are at least two issues that can be investigated further. First, the supermodularity approach proposed in this paper can be used to analyze how other variables (e.g., communication technology and productivity draw) affect firm hierarchy and wages. This approach can also be applied to study properties of other hierarchy models such as the monitoring-based hierarchies (e.g., Calvo and Welllisz 1978, 1979; Qian 1994; Chen 2015). Second, how delay cost affects aggregate productivity and overall wage inequality is also worth exploring. We leave all these interesting questions for future research. From an empirical point of view, better data on internal firm organization (e.g., the number of layers, the span of control) is needed to improve our empirical work. Furthermore, the cost of delay differs substantially across industries due to differences in volatility of demand and competition. Therefore, cross-industry analysis is helpful to make the case that delay cost is indeed important consideration for understanding organizational hierarchy and wage structure inside the firm.

References

Acemoglu, Daron, Philippe Aghion, Claire Lelarge, John Van Reenen, and Fabrizio Zilibotti (2007): "Technology, Information, and the Decentralization of the Firm," *Quarterly Journal of Economics* 122: 1759–1799.

Aoki, Masahiko (1989): "Information, Incentives and Bargaining in the Japanese Economy: A Microtheory of the Japanese Economy," Cambridge University Press.

Aoki, Masahiko (1990): "Toward an Economic Model of the Japanese Firm," *Journal of economic literature* 28: 1–27.

Beggs, Alan W. (2001): "Queues and Hierarchies," *Review of Economic Studies* 68: 297–322.

Bernard, Andrew B., J. Bradford Jensen, Stephen J. Redding, and Peter K. Schott (2007): "Firms in International Trade," *Journal of Economic Perspectives* 21: 105–130.

Bloom, Nicholas, Raffaella Sadun, and J. Van Reenen (2010): "Does Product Market Competition Lead firms to Decentralize?" *American Economic Review* 100: 434–438.

Bolton, Patrick, and Dewatripont, Mathias (1994): "The Firm as a Communication Network," *Quarterly Journal of Economics* 109: 809–839.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt (2002): "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *Quarterly Journal of Economics* 117: 339–376.

Caliendo, Lorenzo, and Esteban Rossi-Hansberg (2012): "The Impact of Trade on Organization and Productivity," *Quarterly Journal of Economics* 127: 1393–1467.

Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg (2015): "The Anatomy of French Production Hierarchies," *Journal of Political Economy* 123(4): 809–852.

Calvo, Guillermo, and Stanislaw Wellisz (1978): "Supervision, Loss of Control, and the Optimum Size of the Firm," *Journal of Political Economy* 86: 943–952.

Calvo, Guillermo, and Stanislaw Wellisz (1979): "Hierarchy, Ability, and Income Distribution," *Journal of Political Economy* 87: 991–1010.

Caroli, Eve, and John Van Reenen (2001): "Skill-Biased Organizational Change? Evidence from a Panel of British and French Establishments," *Quarterly Journal of Economics* 116: 1449–1492.

Cavusgil ,Tamer S., Gary Knight, and John R. Riesenberger (2012): "International Business", 2d. ed., Upper Saddle River, NJ: Pearson Education.

Chen, Cheng (2015): "Management Quality, Firm Organization and International Trade," http://www.sef.hku.hk/~ccfour/MQFOIT.pdf University of Hong Kong, unpublished manuscript.

Djankov, Simeon, Caroline Freund, and Cong S. Pham (2010): "Trading on Time," *Review of Economics and Statistics* 92(1): 166–173.

Fernandes, Ana M. (2007): "Trade Policy, Trade Volumes and Plant-Level Productivity in Colombian Manufacturing Industries," *Journal of International Economics* 71: 52–71.

Frederic, Benjamin (2015): "Trade Shocks, Firm Hierarchies and Wage Inequality," Yale University, unpublished manuscript.

Galbraith, Jay R. (1977): "Organization Design," Reading, Mass.: Addison-Wesley.

Garicano, Luis (2000): "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy* 108: 874–904.

Garicano, Luis, and Esteban Rossi-Hansberg (2004): "Inequality and the Organization of Knowledge," *American Economic Review* 94: 197–202.

Garicano, Luis, and Esteban Rossi-Hansberg (2006): "Organization and Inequality in a Knowledge Economy," *Quarterly Journal of Economics* 121: 1383–1435.

Garicano, Luis, and Esteban Rossi-Hansberg (2012): "Organizing Growth," *Journal of Economic Theory* 147: 623–656.

Garicano, Luis, and Van Zandt (2012): "Hierarchies and the Division of Labor" in *The Handbook of Organizational Economics*, edited by Robert Gibbons and John Roberts, Princeton University Press.

Guadalupe, Maria, and Julie M. Wulf (2010): "The Flattening Firm and Product Market Competition: The Effect of Trade Liberalization on Corporate Hierarchies," *American Economic Journal: Applied Economics* 2: 105–127.

Melitz, M. J. (2003): "The Impact of Trade on Intra-industry Reallocations and Aggregate Industry Productivity," *Econometrica* 71(6): 1695–1725.

Patacconi, Andrea (2009): "Coordination and Delay in Hierarchies," *RAND Journal of Economics* 40: 190–208.

Qian, Yingyi (1994): "Incentives and Loss of Control in an Optimal Hierarchy," *Review of Economic Studies* 61: 527–544.

Radner, Roy (1992): "Hierarchy: The Economics of Management," *Journal of Economic Literature* 30: 1382–1415.

Radner, Roy (1993): "The Organization of Decentralized Information Processing," *Econometrica* 61: 1109–1146.

Rajan, Raghuram G., and Julie M. Wulf (2006): "The Flattening Firm: Evidence on the Changing Nature of Firm Hierarchies from Panel Data," *Review of Economics and Statistics* 88: 759–773.

Roberts, Mark J. (1996): "Colombia, 1977–1985: Producer Turnover, Margins, and Trade Exposure," pp. 227–259 in *Industrial Evolution in Developing Countries: Micro Patterns of Turnover, Productivity, and Market Structure*, edited by Mark Roberts and James Tybout. Oxford: Oxford University Press.

Roberts, Mark J. and James R. Tybout (1997): "The Decision to Export in Colombia: An Empirical Model of Entry with Sunk Costs," *American Economic Review* 87: 545–564.

Sabur, S. A., M. S. Palash, M. H. Rahman, and M. Miah (2004): "Profitability And Problems Of Exporting Fresh Vegetable From Bangladesh," *Bangladesh Journal of Agricultural Economics* 27: 99–106.

Tag, Joacim (2013): "Production Hierarchies in Sweden," *Economics Letters* 121: 210–213.

Topkis, Donald M. (1998): "Supermodularity and Complementarity," Princeton, NJ: Princeton University Press.

Whitney, Daniel (1988): "Manufacturing by Design," *Harvard Business Review* 66(4): 83–91.

A. Appendix

A.1. Summary Statistics

Table 2. *Summary Statistics for Production Information*

	value of production	employment	value added	book value of fixed assets
no. obs.	76083	76083	69285	76094
mean	629962.9	69.51	290255.5	114154.9

Unit of currency is Colombian peso.

Table 3. *Summary Statistics for Organization Structure*

Variable	no. obs.	mean	std. dev.	min	max
layers	76083	2.512	0.637	0	3
layers-2	76083	2.846	0.916	0	4
span	74863	5.661	13.10	0	625
span-2	74863	5.041	12.06	0	625

Table 4. *Summary Statistics for Average Wage at Different Levels*

Variable	no. obs.	mean	std. dev.	min	max
managers	47384	2394.6	3706.3	2	60691
technicians	26059	1528.1	1652.9	34.5	17159.7
skilled workers	67876	796.9	825.7	29	11361.2
unskilled workers	74275	576.5	545.4	20	7408.0

Top and bottom 0.2% observations (potential outliers) are excluded.

Table 5. *Summary Statistics for Employment at Each Level*

Variable	no. obs.	mean	std. dev.	min	max
number of owners	76083	1.142	16.72	0	3693
number of managers	76083	2.000	16.77	0	3693
number of technicians	76083	2.239	12.52	0	986
number of skilled workers	76083	15.76	49.16	0	1563
number of unskilled workers	76083	48.37	121.9	0	5174

A.2. Transition in the Number of Layers

In this subsection, we document how Colombian plants had changed their number of layers empirically. First, we report the fraction of plants that had changed the number of layers from year $t-1$ to year t among plants that existed in both year t and year $t-1$ (i.e., we exclude firms that entered in year t or exited in year $t-1$):

Table 6. *Fraction of Plants Changing the Number of Layers from Year $t-1$ to Year t*

	layers	*layers-2*
Fraction	$10.43\%(=6408/61447)$	$16.27\%(=9955/61189)$

From Table 6, we notice that a non-negligible fraction of plants had changed the number of layers. Furthermore, fractions reported in Table 6 are lower bounds for two reasons. First, reorganization through adding or dropping layers is a big change for a firm which might take several years to be completed. Second, our measure of the number of layers is coarse. Next, we report the distribution of the numbers of layers for Colombian plants. From Table 7, we find that it is not true that most plants had three or four layers. In fact, Table 8 shows that many plants had skilled and unskilled workers. However, a substantial fraction of plants did not have technicians and managers.

Table 7. *Distribution of the Number of Layers*

	layers $=1$	*layers* $=2$	*layers* $=3$	
rel. freq.	5.9%	35.4%	58.8%	
	layers-2 $=1$	*layers*-2 $=2$	*layers*-2 $=3$	*layers*-2 $=4$
rel. freq.	5.8%	29.8%	36.5%	27.9%

Table 8. *Distribution of the Number of Workers*

Layer of	Unskilled Workers	Skilled Workers	Technicians	Managers
Number of Plants	74585	68161	26179	47630

The definition for the number of layers is *layers-2*.

Third, we report the transition matrix for the change in the number of layers. Tables 9 and 10 show that when Colombian plants changed the number of layers, they usually added or deleted one layer.

Fourth, we explore how plants that had changed the number of layers implemented such a change. We use the definition of *layers-2* here. For 2024 plants that had decreased the number of layers from four to three, we have summary statistics in Table 11. Note that when we say plants, it means a plant in a given year (i.e., plant-year pairs).

Table 9. *Transition Matrix for the Change in the Number of Layers for layers*

Number of Plants	$layers_t = 1$	$layers_t = 2$	$layers_t = 3$
$layers_{t-1} = 1$	2299	611	92
$layers_{t-1} = 2$	514	17951	2745
$layers_{t-1} = 3$	109	2205	34571

Table 10. *Transition Matrix for the Change in the Number of Layers for layers-2*

Number of Plants	$layers\text{-}2_t = 1$	$layers\text{-}2_t = 2$	$layers\text{-}2_t = 3$	$layers\text{-}2_t = 4$
$layers\text{-}2_{t-1} = 1$	2284	577	101	11
$layers\text{-}2_{t-1} = 2$	486	14655	2268	253
$layers\text{-}2_{t-1} = 3$	101	1897	18311	2126
$layers\text{-}2_{t-1} = 4$	14	223	2024	15766

Table 11. *Transition Pattern for Four-layer Plants*

Change from	4,3,2,1 to 3,2,1	4,3,2,1 to 4,2,1	4,3,2,1 to 4,3,1	4,3,2,1 to 4,3,2
Number of Plants	337	1590	65	32

The definition for the number of layers is *layers-2*. Layer 4: managers; Layer 3: technicians; Layer 2: skilled workers; Layer 1: unskilled workers.

Table 12. *Transition Pattern for Three-layer Plants*

Change from	3,2,1 to 4,3,2,1	4,2,1 to 4,3,2,1	4,3,1 to 4,3,2,1	4,3,2 to 4,3,2,1
Number of Plants	431	1563	110	22
Change from	4,2,1 to 2,1	4,2,1 to 4,1	3,2,1 to 3,1	3,2,1 to 2,1
Number of Plants	1090	251	21	387

The definition for the number of layers is *layers-2*. Layer 4: managers; Layer 3: technicians; Layer 2: skilled workers; Layer 1: unskilled workers. We omit other patterns (e.g., 4,3,2 to 3,2), since there are too few observations that had such changes.

Table 13. *Transition Pattern for Two-layer Plants*

Change from	2,1 to 3,2,1	2,1 to 4,2,1	4,1 to 4,2,1	3,1 to 3,2,1
Number of Plants	341	1393	376	29
Change from	4,1 to 1	2,1 to 2	2,1 to 1	3,1 to 1
Number of Plants	75	50	311	16

The definition for the number of layers is *layers-2*. Layer 4: managers; Layer 3: technicians; Layer 2: skilled workers; Layer 1: unskilled workers. We omit other patterns (e.g., 4,1 to 4,3,1, or 4,3 to 3), since there are too few observations that had such changes.

For 577 plants that decreased the number of layers from one to two, we have summary statistics in Table 14.

Table 14. *Transition Pattern for One-layer Plants*

Change from	1 to 1, 4	1 to 1, 3	1 to 1, 2	2 to 2, 1
Number of Plants	79	18	438	24

The definition for the number of layers is *layers*-2. Layer 4: managers; Layer 3: technicians; Layer 2: skilled workers; Layer 1: unskilled workers. We omit other patterns (e.g., 2 to 2, 3), since there are too few observations that had such changes.

Tables 11 to 14 show that the layer of managers and that of technicians are the ones that are most likely to be added and dropped when Colombian plants reorganized.

Fifth, we show how average wage and employment change when plants adjust the number of layers. Table 15 shows that when Colombian plants increased (or decreased) the number of layers, employment increased (or decreased) substantially. Interesting, this pattern does not hold for the change in average deflated wage, as Table 16 shows.

Table 15. *Average Change in Employment when Plants Change the Number of Layers*

Type	Plants that delayer	Plants that increase the Number of Layers
Average Change in Employment	-5.71	4.29
No. Obs.	4921	5410

The definition for the number of layers is *layers*-2. Median and average employment of all observations are 26 and 69.5 respectively.

Table 16. *Average Change in Deflated Wage when Plants Change the Number of Layers*

Type	Plants that delayer	Plants that increase the Number of Layers
Average Change in deflated wage	0.15	1.74
No. Obs.	4902	5410

The definition for the number of layers is *layers*-2. Median and average deflated wage of all observations are 17.3 and 23.2.

Finally, we show that how exporting status is related the change in the number of layers. First, As Table 17 shows, the correlation between exporting status and the number of layers is positive, if we do not control for the firm fixed effects and other firm-level variables. However, when we exploit cross-time variation in exporting status, the correlation turns out to be insignificant, as Table 18 shows. This shows that positive correlation between exporting status and the number of layers is mainly due to the firm fixed effects and firm size. Next, we show how the number of layers, average wage and average employment changed, when Colombian plants started or stopped exporting. Table 19 shows that, compared with plants that did not change exporting status, both exporting starters and exporting stoppers decreased the number

of layers. The first finding (for exporting starters) is what our theory predicts, while the second one (for exporting stoppers) is mainly due to the substantial reduction in firm size. Tables 20 and 21 demonstrate that compared with plants that did not change exporting status, exporting starters did not increase employment and wage (on average) substantially, while exporting stoppers reduced their employment substantially.

Table 17. *Correlation between Exporting Status and the Number of Layers at the Cross-Sectional Level*

	layers	*layers*-2
Correlation Coefficient	0.1711	0.2509

Correlation coefficients are statistically significant at 1% level. Number of observations: 76083.

Table 18. *Correlation between the Change in Exporting Status and the Change in the Number of Layers (and Average Wage and Employment)*

	Δ*layers*	Δ*layers*-2	Δ*Ave. Employment*	Δ*Ave. Wage*
Correlation Coefficient	0.0016	0.0054	0.0097*	0.0004

$*$:Correlation coefficients are statistically significant at 10% level. Number of observations: 61573.

Table 19. *Correlation between the Change in the Exporting Status and the Change in the Number of Layers*

	No. Obs.	Δ*layers*	Δ*layers*-2
Exporting Starters in year t	1583	0.0025	0.0057
Exporting Stoppers in year t	1040	-0.007	-0.026
Other Plants	58950	0.0071	0.0074

In short, the substantial reduction in employment for exporting stoppers can be used to explain why these plants reduced the number of layers (in the relative sense). For exporting starters, the (small) increase in employment should imply that they increase the number of layers after beginning to export. However, Table 19 shows the opposite pattern, which leaves room to our story of delay cost to explain this finding.

Table 20. *Correlation between the Change in the Exporting Status and the Change in Average Employment*

	No. Obs.	Δ*Average Employment*
Exporting Starters in year t	1566	1.21
Exporting Stoppers in year t	1032	-4.13
Other Plants	58727	-0.08

Top and bottom 0.2% observations (potential outliers) are excluded. Median and average employment of all observations are 26 and 69.5 respectively.

Table 21. *Correlation between the Change in the Exporting Status and the Change in Average Wage*

	No. Obs.	Δ*Average Wage*
Exporting Starters in year t	1557	1.20
Exporting Stoppers in year t	1025	0.55
Other Plants	58690	0.58

Top and bottom 0.2% observations (potential outliers) are excluded. Median and average deflated wage of all observations are 17.3 and 23.2.