Management Quality, Firm Organization and International Trade*

Cheng Chen†
Department of Economics
Princeton University
Job Market Paper
The Latest Version: http://www.princeton.edu/~ccfour/JMP.pdf
Online Appendix: http://www.princeton.edu/~ccfour/JMPonline.pdf

November 22, 2013

Abstract

The quality of management technology that is used to monitor and incentivize workers varies substantially across countries. To understand the impact of this on economic activities, I develop a two-sector model in which firms facing heterogeneous demands set up hierarchies to manage the production processes in a monopolistically competitive sector. Entrepreneurs decide the number of hierarchical layers, the effort level of each worker, and the span of control of supervisors. I then use the theory to explain two empirical findings established in the literature. First, a common improvement in this type of management technology across all firms intensifies competition in the monopolistically competitive sector. As a result, the smallest firms are forced to leave the market; the most efficient firms thrive; and the average firm size increases. Second, firms are less decentralized in economies with ineffective management technology. In an extended two-country model incorporating international trade, I show that firms facing increasing import competition flatten their hierarchies and use more incentive-based pay. Furthermore, I find that countries with superior management technology experience larger welfare gains from opening up to trade.

Key words: Management; Firm organization and productivity; Trade liberalization; Institutions and development
JEL Classification: D21; D23; F12; L22; L23; O1; O4

*I am grateful to Gene Grossman, Stephen Redding, and Esteban Rossi-Hansberg for invaluable guidance. I also thank Pol Antràs, Mark Aguair, Alexis Antoniades, Costas Arkolakis, Davin Chor, Liang Dai, Wouter Dessein, Taiji Furusawa, Hideshi Itoh, Oleg Itskhoki, Greg Kaplan, Nobuhiro Kiyotaki, Kala Krishna, John Mclaren, Guido Menzio, Eduardo Morales, Benjamin Moll, Richard Rogerson, Felix Tintelnot, Sharon Traiberman, John Van Reenen, Vaidyanathan Venkateswaran, Jonathan Vogel, Hongsong Zhang, Ruilin Zhou, and numerous seminar participants for their valuable comments. Financial support from the International Economics Section at Princeton University is greatly appreciated.

†Department of Economics, Princeton University, Princeton, NJ 08544, USA. Homepage: http://scholar.princeton.edu/ccfour. E-mail: ccfour@princeton.edu.

1 Introduction

Recent empirical research using firm-level survey data from various countries has substantiated the existence of large variation in the quality of management technologies across countries.[1] Furthermore, the quality of management technologies has been shown to have considerable impact on firm performance and firm organization (Bloom and Van Reenen (2007, 2010), and Bloom, Sadun, and Van Reenen (2012a)). However, relatively little is known about the aggregate implications of differences in management technologies. More specifically, how does the quality of management technologies affect the firm size distribution, organizational structures of firms with different efficiency levels and average productivity of firms in an economy? To study these questions, I develop a general equilibrium model of heterogeneous firms in differentiated product markets that incorporates one type of management technology and *endogenous* managerial organization. I focus on the canonical approach to modeling endogenous firm organization based on effort and incentives, which have been empirically shown to be important dimensions of management practice. I show that a management technology that allows firms to better monitor and incentivize employees amplifies the welfare gains from trade and generates a pro-competitive effect that facilitates resource reallocation from less efficient firms to more efficient firms. As a result, firms are bigger and more decentralized on average, and average productivity of firms increases when the management technology improves.

This paper focuses on the quality of a particular type of management technology: the ability to monitor and incentivize employees given a firm's organizational choices. From now on, for the sake of simplicity I use management technology (MT) to denote management technology used to monitor and incentivize employees. I focus on this dimension of MT because it is an important component of overall management technology and affects firm performance substantially.[2] This type of MT is soft technology that consists of various management rules. Management rules that either specify regular performance tracking and review or remove poor performers help a firm monitor workers and punish shirking employees.

Large differences in the quality of MT are beyond the control of the firm. For instance, low-quality institutions such as rigid labor markets and weak law enforcement negatively affect the ability of firms to punish misbehaving employees (e.g., Bloom and Van Reenen (2010) and Bloom et al. (2013)). Moreover, better management rules diffuse slowly across borders and do not exist in many countries because of information barriers.[3] Hence, I treat MT as exogenous from the perspective of a firm, but allow firms to make endogenous choices about management organization subject to this technology.

The economic objects I want to analyze interact together, and MT seems to play a role in determining them. First, differences in the firm size distribution across countries have implications for resource misallocation and aggregate productivity (Hsieh and Klenow

[1] For a discussion of management as a technology, see Bloom, Sadun, and Van Reenen (2012b). management technologies defined in this paper are the same as management practices defined in Bloom and Van Reenen (2007, 2010). Examples include good management rules to remove poor performers and check employees' behavior effectively.

[2] For details on the overall management quality and effects of monitoring and incentives on firm performance, see Appendix 7.1.

[3] Bloom et al. (2013) point out that one major reason why Indian firms are poorly managed is that their managers do not know about the existence of better management technologies.

(2009)). Second, the organizational structure of firms matters for firm performance and intra-firm wage inequality (Caliendo, Monte, and Rossi-Hansberg, 2012). Most importantly, all of these are systematically related to MT. For instance, Bloom et al. (2013) argue that one major reason for why efficient firms can't expand fast in India is that they are unwilling to decentralize the production processes due to bad MT. Because of the slow expansion of efficient firms, many small and inefficient firms survive in India, which is one of the reasons why aggregate productivity of firms is low in India. In summary, MT is a candidate to explain differences in aggregate-level and firm-level outcomes across economies.

This paper develops a general equilibrium model with two sectors. One sector is a homogeneous sector. It is a perfectly competitive sector with a constant returns to scale technology producing a homogeneous good. I assume that there are no monitoring and incentive issues inside firms of this sector. The sector that is the main focus of my analysis is the monopolistically competitive sector, which comprises a continuum of differentiated products with a constant elasticity of substitution (CES) à la Dixit and Stiglitz (1977). The purpose of having the homogeneous sector is to endogenize the expected wage of workers in the CES sector in a tractable way. The demand for these products varies depending on their individual characteristics. An entrepreneur can enter this sector by paying a fixed cost, and then she receives a random draw of demand (or quality) for her product. The demand draw and the quality draw are isomorphic in this framework, hence I will refer to them interchangeably. Once the entrepreneur observes the quality, she decides whether or not to stay in the market as there is a fixed cost to produce as well. In equilibrium, entrepreneurs in the monopolistically competitive sector earn an expected payoff that is equal to their exogenous outside option due to free entry.[4] For simplicity I refer to the monopolistically competitive sector henceforth as the CES sector.

Firms in the CES sector need to monitor and incentivize employees, as production requires both time and effort, and the latter is costly for firms to observe. Following the canonical approach to modeling monitoring and incentive problems within a firm (i.e., Calvo and Wellisz (1978, 1979) and Qian (1994)), I assume that the firm sets up a hierarchy to monitor workers and provide incentives. A hierarchy is an organization with multiple layers, and a layer is a group of workers who have the same level of seniority. More specifically, the firm allocates workers into different layers to make supervisors monitor their direct subordinates and offer incentive-compatible wage contracts to workers. In equilibrium, production workers (i.e., workers in the bottom layer) and non-production workers are incentivized to exert effort to produce output and monitor subordinates respectively.

Firms whose products have greater demand set up a hierarchy with more layers. In addition to output and price, firms choose the number of layers as well as the span of control at each layer. The span of control is defined as the ratio of the number of supervisors to the number of their direct subordinates. When the firm wants to produce more, it has to increase the span of control owing to the constraint of managerial talent at the top. A larger span of control implies that less attention is paid to monitor each subordinate, which has to be compensated by higher wage since the firm needs to prevent workers from shirking. As a result, the marginal cost (MC) increases. The firm can add a layer and decrease the span of control to save wage payments to workers at existing layers,

[4]The expected *payoff* equals expected profit minus the disutility to exert effort.

which makes the MC drop. However, this comes at the cost of extra wage payments to workers at the new layer. In short, adding a layer is like an efficiency-enhancing investment with a fixed cost. Firms whose goods are more preferred by consumers have an incentive to set up a hierarchy with more layers, as they produce more in equilibrium.

In order to study the pro-competitive effect of improved MT, I consider a scenario in which the quality of MT, which is common across firms, improves. This occurs when the labor market is deregulated or better management rules are introduced into an economy. Such an improvement benefits all firms by reducing their labor costs. However, this benefit is *uneven* across firms. Firms with more layers gain disproportionately more, as their average variable costs (AVCs) increase less rapidly with output. As a result, firms with the worst demand draws are forced to leave the market, and firms with the best demand draws thrive.[5] Furthermore, the resulting firm size distribution and the distribution of the number of layers move to the right in the first-order-stochastic-dominance (FOSD) sense. In other words, firms are bigger and "taller" on average in economies with superior MT. These results are consistent with two findings from Hsieh and Klenow (2009, 2012). First, China and India whose firms have lower management scores than U.S. firms have many small firms and a lower average firm size. Second, the U.S. whose firms have high management scores are better at getting efficient firms to grow than are China and India. Third, all surviving firms either increase the number of layers or make the span of control larger after an improvement in MT. Therefore, firms are more decentralized in economies with superior MT, which is consistent with the findings in Bloom, Sadun, and Van Reenen (2012a) and Bloom et al. (2013). In summary, a common improvement in MT across all firms generates a pro-competitive force that reallocates resources toward more efficient firms.

Aggregate productivity of firms increases as a result of an improvement in MT. Following Caliendo and Rossi-Hansberg (2012), I use the inverse of unit costs to measure firm productivity. Following Olley and Pakes (1996), I use the weighted average of firm productivity to measure aggregate productivity. Gains in the weighted average of firm productivity come from three sources. First, the least productive firms exit the market after an improvement in MT. Second, market shares of the most productive firms increase after an improvement in MT. Finally, productivity of surviving firms increases, as improved MT reduces firm costs. In total, these three effects together increase the weighted average of firm productivity.[6]

Beyond macro-level implications, the model has micro-level predictions as well. First, although wages at all layers increase with firm size given the number of layers, they fall

[5] In a hypothetical world, if firms were forced to have the same number of layers, this uneven effect across firms and the exit of the smallest firms would disappear after MT improves. This is because all firms have the same AVC function. Therefore, endogenous selection into the hierarchy with different numbers of layers is the key to generating a differential impact of an improvement in MT on firms with different demand draws.

[6] Powell (2013) investigates how contract enforcement affects the *dispersion* of the firm productivity distribution in a perfectly competitive product market. One key result from the paper is that weak enforcement of laws hurts unproductive firms more due to the existence of a fixed production cost and a dynamic-enforcement constraint. As a result, the distribution of firm productivity is more dispersed in economies with weaker enforcement of laws. Both Powell (2013) and my paper emphasize the role of institutions in shaping aggregate productivity. Powell (2013) focuses on the second-order moment of the distribution of firm productivity, while my paper focuses on the first-order moment of it.

at existing layers when the firm adds a layer. This reduction occurs because the addition of a new layer decreases the span of control at existing layers. This result shows that employees might lose from firm expansion. Second, when wages increase, they increase disproportionately more at upper layers, which leads to a bigger wage ratio between two adjacent layers. Similarly, when wages fall, they fall disproportionately more at upper layers, which leads to a smaller wage ratio between two adjacent layers. These results imply a distributional effect on workers' wages due to firm expansion. Third, in the theory, firms that are bigger or more efficient have more layers. This is because adding a layer is like an investment that requires a fixed cost and reduces the firm's MC. In total, all the above results on firm-level outcomes are consistent with the evidence presented in Caliendo, Monte, and Rossi-Hansberg (2012) and have implications for intra- and inter-firm wage inequality.

Most countries are open economies, and trade liberalization brings changes to welfare and management practices of firms. I extend the baseline model into the international context à la Melitz (2003) to discuss how trade liberalization and firm management interact with each other.[7] Two interesting results emerge. At the firm level, the internal organization of firms changes after trade liberalization. More specifically, non-exporting firms flatten their hierarchies by reducing the number of layers and increasing the span of control. Furthermore, non-exporters increase the amount of incentive-based pay after import competition has increased. Both results are consistent with the findings from Guadalupe and Wulf (2010) that American firms facing increasing import competition from Canada flattened their hierarchies and used more incentive-based pay. At the country level, the quality of MT affects the welfare gains from trade. I find that an economy with superior MT benefits disproportionately more from opening up to trade. This result suggests that countries with poor institutions that hamper firm management are likely to gain less from participation in the world economy.[8]

This paper contributes to the literature on incentive-based hierarchies in three ways.[9] First, I treat the number of layers as a discrete variable and solve the optimal number of layers by characterizing firm's cost functions.[10] Treating the number of layers as a discrete variable is empirically realistic and important for the model's predictions on wages and firm productivity. Second, firm size is endogenously determined, as each firm faces a downward-sloping demand curve.[11] Finally, I incorporate the canonical model of incentive-based hierarchies into a general equilibrium setting. By doing so, I can analyze the impact of improved MT on the firm size distribution and average productivity of

[7]Papers that incorporate the monitoring-based wage determination into an international trade model include those by Copeland (1989), Matusz (1996), Chen (2011), and Davis and Harrigan (2011).

[8]In a different setting, Kambourov (2009) finds that institutional features of labor markets affect responses of an economy to trade liberalization. Particularly, higher costs to fire workers dampen the welfare gains from trade.

[9]Earlier research papers on management hierarchies include those by Williamson (1967), Beckmann (1977), and Keren and Levhari (1979) etc. For more details of research on incentive-based hierarchies, see Mookherjee (2010).

[10]Calvo and Wellisz (1979) do not endogenize the number of layers, while Qian (1994) treats the number of layers as a continuous variable.

[11]Calvo and Wellisz (1978) point out that firm size is undetermined in the standard model of incentive-based hierarchies. Firm size in Qian (1994) is exogenously given in some sense, as the firm is assumed to have a fixed amount of capital and a fixed capital-labor ratio to produce. If the firm is allowed to choose the capital level optimally, firm size goes to infinite, as pointed out by Meagher (2003).

firms, which are general equilibrium objects. Furthermore, these implications can be readily contrasted with the data and help explain the stylized patterns observed in the real world.

The literature on knowledge-based hierarchies (e.g., Garicano (2000), Garicano and Rossi-Hansberg (2004, 2006, 2012), Caliendo and Rossi-Hansberg (2012)) has been successful at providing a framework to analyze how the information and communication technology (ICT) affects various economic outcomes such as wage inequality and economic growth. However, it is silent on the role of MT in determining firm-level outcomes and aggregate economic variables such as the firm size distribution and aggregate productivity. A paper in this literature that is closely related to mine is Caliendo and Rossi-Hansberg (2012). There are two major differences between these two papers. First, the moral hazard problem that is absent in Caliendo and Rossi-Hansberg (2012) is the key ingredient of this paper. I focus on a different mechanism (effort and incentives), and it yields some different predictions such as what matters for the firm size distribution is not the ICT but institutions that affect the ability of firms to monitor and incentivize employees. Second, this paper yields some important micro-level predictions that Caliendo and Rossi-Hansberg (2012) do not have such as predictions on relative wages.

This article is related to the literature on heterogeneous firms and international trade (e.g., Bernard, Eaton, Jensen and Kortum (2003), Melitz (2003), Yeaple (2005) and Melitz and Ottaviano (2007)). This paper complements the literature on heterogeneous firms and international trade by showing how organizational structures of firms and the quality of MT affect the responses of firms and an economy to trade liberalization. In particular, the welfare gains from trade is impacted by the quality of MT which is affected by institutional quality.

The remainder of the paper is organized as follows. Section two solves both the individual firm's optimization problem and the problem of resource allocation in general equilibrium. Section three investigates how differences in MT across economies (and firms) affect various economic activities. Section four extends the baseline model to include international trade and explores how trade liberalization affects firms' internal organization and welfare. Section five presents evidence to support the main theoretical predictions derived in Section three. Section six concludes.

2 The Model

In this section, I develop a model of the hierarchical firm that features firms' endogenous selection of a hierarchy with a specific number of layers. The key elements are the firm's decisions on the span of control as well as the number of layers. I will subsequently introduce the model into a general equilibrium setting and solve the problem of resource allocation in both the product and labor markets.

2.1 Environment

The economy comprises two sectors, L units of labor and N potential entrepreneurs, where N is sufficiently large that the free entry (FE) condition discussed below will hold with equality. One sector produces a homogeneous good and is perfectly competitive, while

the other sector produces horizontally differentiated goods and features monopolistic competition.

A representative agent demands goods from both sectors and has the following Cobb-Douglas utility function:

$$U = \Big(\frac{C_c}{\gamma}\Big)^{\gamma}\Big(\frac{C_h}{1-\gamma}\Big)^{1-\gamma} - I\psi(a_i), \tag{1}$$

where C_h is the consumption of the homogeneous good and C_c is an index of consumption of differentiated goods defined as

$$C_c = \left(\int_{\Omega} \theta^{\frac{1}{\sigma}} y(\theta)^{\frac{\sigma-1}{\sigma}} M\mu(\theta)d\theta\right)^{\frac{\sigma}{\sigma-1}}, \tag{2}$$

$y(\theta)$ is the consumption of variety θ, M denotes the mass of products available to the consumer, $\mu(\theta)$ indicates the probability distribution over the available varieties in Ω, and $\sigma > 1$ is the constant elasticity of substitution. Note that θ is a demand shifter for a firm variety, so agents demand more of goods with higher θ at a given price. I and $\psi(a_i)$ are, respectively, an indicator function and a disutility to exert effort that will be discussed later. The final composite good is defined as

$$\Big(\frac{C_c}{\gamma}\Big)^{\gamma}\Big(\frac{C_h}{1-\gamma}\Big)^{1-\gamma},$$

which is the first part of terms appearing in the right hand side of equation (1). I choose the price of it as the numeraire, so

$$P^{\gamma} p_h^{1-\gamma} \equiv 1, \tag{3}$$

where

$$P = \Big(\int_{\Omega} \theta p(\theta)^{1-\sigma} M\mu(\theta)d\theta\Big)^{\frac{1}{1-\sigma}} \tag{4}$$

is the ideal price index of the differentiated goods. p_h is the price of the homogeneous good, and $p(\theta)$ is the price of variety θ.

The homogeneous sector features no frictions, and the perfectly competitive market structure implies that firms receive zero profit. Labor is the only factor used in production, and the production technology implies that output equals the number of workers (i.e., labor) employed. The price of the homogeneous good is also the wage offered in this sector. There is no unemployment among workers who enter this sector in equilibrium owing to the absence of frictions.

The CES sector produces a continuum of differentiated products. The demand for these products varies depending on their individual characteristics. There is a large pool of potential entrepreneurs who have managerial ability to set up firms in this sector. An entrepreneur can enter this sector and receive a random draw of quality for her product after paying a fixed cost f_1 to design it. Given the existence of a fixed cost f_0 to produce, the entrepreneur decides whether or not to stay in the market after she observes her quality draw. Both the entry cost and the fixed cost are paid in the form of the final

composite good, as in Atkeson and Burstein (2010). The entrepreneur has to employ workers and organize the production process if she decides to produce.

Workers choose the sector in which they seek employment, while entrepreneurs choose whether or not to operate a firm. Both types of agents are risk neutral. In equilibrium, workers' expected payoff from entering both sectors must be the same since they can freely move between sectors. I assume that the outside option (or reservation utility) of an entrepreneur is forgone, if she chooses to enter the CES sector. Thus, the expected payoff of entrepreneurs who choose to enter the CES sector equals their exogenous outside option f_2 due to the free entry of firms in equilibrium.[12] Workers cannot choose to be entrepreneurs, as they don't have managerial talents. Furthermore, I assume that f_2 is big enough that the expected payoff of workers is strictly smaller than it in equilibrium. Therefore, entrepreneurs have no incentives to become workers.

2.2 The Organization of Production

I follow the literature on incentive-based hierarchies (e.g., Calvo and Wellisz (1978, 1979)) in modeling the organization of production. More specifically, I assume that each firm has to employ workers at various layers and incentivize them to exert effort in order to produce. Production workers only produce output, while non-production workers only monitor and incentivize their direct subordinates.[13]

Production requires effort and time of workers. The worker's effort choice a_i is assumed to be a binary variable between working and shirking (i.e., $a_i \in \{0, 1\}$) for reasons of tractability.[14] The input of workers' time equals the number of workers. Production workers produce output, and shirking results in defective output that cannot be sold. Thus, the production function is

$$q = \int_0^{m_T} a(j)dj, \tag{5}$$

where m_T is the number of production workers and $a(j)$ is the effort level of the j-th unit of labor inputs. Here I assume that labor inputs are divisible, as workers' time is

[12] Following Melitz (2003), I treat the outside option of entrepreneurs as exogenous.

[13] Most firms monitor and incentivize their employees in reality. For real-world examples, see `http://matthewoudendyk.blogspot.com/2006/11/how-monitoring-is-being-do_116260155005227748.html` and `http://management.about.com/cs/people/a/MonitorEE062501.htm`. Management activities used to monitor and incentivize employees include a variety of work done by supervisors. First, supervisors monitor their subordinates using information technology such as video surveillance, e-mail scanning and location monitoring (Hubbard (2000, 2003)). Second, monitoring also happens when supervisors communicate with their subordinates and try to check whether or not the subordinates are working hard. Finally, business meetings in which supervisors evaluate subordinates' performance and decide whether or not to reward good performers and fire poor performers are important parts of monitoring and incentivizing activities as well. Admittedly, monitoring and incentivizing subordinates are parts of what non-production workers do in reality. I focus on this dimension of non-production workers' work in order to emphasize the impact of improvements in MT to monitor and incentivize employees on economic outcomes. In an ongoing work, I consider a model in which non-production workers both monitor their subordinates and contribute to production by solving problems raised by the subordinates.

[14] As shown in the online appendix, it is straightforward to generalize the analysis to allow effort to be a continuous variable.

divisible. Non-production workers at layer i monitor their direct subordinates at layer $i+1$ and need to be monitored by supervisors at layer $i-1$ as well. A smaller i denotes a higher layer in the firm's hierarchy, and the entrepreneur is at layer zero. Layer T is the lowest layer in the hierarchy which is occupied by production workers. Thus, production workers and non-production workers have different roles in the production process.

Workers must be monitored if the firm wants them to exert effort. The firm has no reason to fire a shirking worker unless it is able to detect his misbehavior. A worker at layer i is induced to work for wage w_i, if and only if

$$w_i - \psi \geq (1 - p_i)w_i, \tag{6}$$

where $p_i(\leq 1)$ is the probability of catching and firing a shirking worker, and ψ is the disutility of exerting effort. A worker's utility differs from the utility of consuming goods only when he works in the CES sector and exerts effort in the production process (i.e., $I = 1$ in equation (1)). The above inequality is the incentive compatibility constraint that the payoff obtained from exerting effort must be greater than or equal to that of shirking.

The probability of catching and firing a shirking worker depends on two factors: the adjusted span of control and MT. First, the bigger the adjusted span of control, the less frequently a subordinate's behavior is checked by his supervisor (less time or monitoring effort is spent on him). This implies a lower probability of *catching* a shirking worker. Second, the quality of MT affects the probability of *detecting* workers' misbehavior. Management rules that clarify performance measures for employees lead to easier detection of workers' misbehavior. Finally, the quality of MT also affects the probability of successfully *firing* shirking workers. Firms that are located in economies with either rigid labor markets or weak law enforcement are found to be worse at using good management rules to remove poor performers (Bloom and Van Reenen (2010)). I capture these effects by assuming the following functional form for $p(b, x_i)$:

$$p(b, x_i) = \frac{1}{bx_i(\theta)}, \tag{7}$$

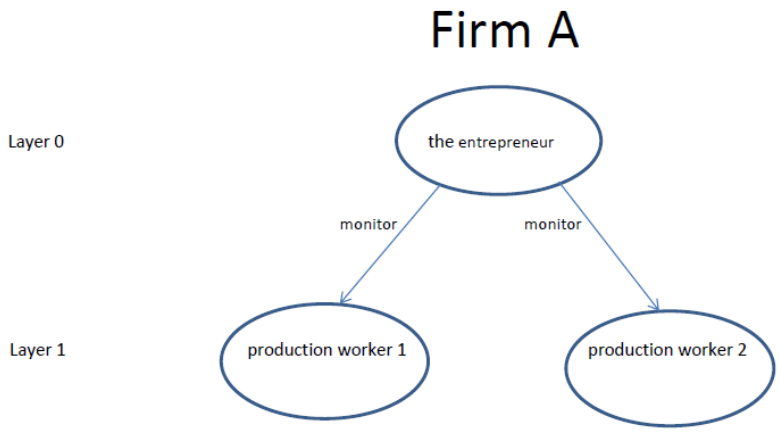
where $x_i(\theta) \equiv \frac{m_i(\theta)}{\int_0^{m_{i-1}} a(j)dj}$ is the span of control adjusted by supervisors' efforts.[15] Parameter b reflects the *inefficiency* of MT. More specifically, the worse the MT is, the bigger the value of b.

A firm may want to hire non-production workers, since it wants to economize on the cost to incentivize workers. I use Figures 1 and 2 to elicit the economic forces behind this choice.[16] Consider firm A that receives a low demand draw θ_A and wants to produce two units of goods as illustrated in Figure 1. The span of control of the entrepreneur is small, which implies the low incentive-compatible wage (i.e., equation (7)) paid to production workers. Thus, it is optimal to have production workers only, as non-production workers do not produce output. Next, consider firm B that receives a high demand draw θ_B

[15] As what I will show, the entrepreneur allocates the monitoring intensity evenly across workers at the same layer. A more flexible functional form is $p(b, x_i) = \frac{1}{bx_i(\theta)^\nu}$ where ν can be different from one. Allowing ν to differ from one does not affect qualitative results of the paper. Detailed discussions are available upon request.

[16] These two figures only serve for illustrative purposes.

Figure 1: A Firm with Two Layers



and wants to produce six units of goods which is illustrated in Figure 2. The incentive-compatible wage paid to production workers would be too high, if the firm did not hire non-production workers between the entrepreneur and production workers.[17] If the firm hires non-production workers who monitor production workers, the incentive-compatible wage paid to production workers will be reduced, which makes the labor costs lower. Obviously, this comes at the cost of extra wage payment to non-production workers. Therefore, it is optimal for the firm to add non-production workers only when the output level is high, which will be shown rigourously in Subsection 2.3. The above logic also explains why the firm wants have two (three, four...) layers of non-production workers when the output level is high. In total, the trade-off between lower wage paid to existing workers and extra wage payment paid to newly employed non-production workers shapes the optimal choice of the number of layers for a firm.

I characterize two optimal choices of the firm before solving the firm's optimal decisions on the other variables (e.g., output and employment etc.), as these two choices are independent of the firm's decisions on the other variables. In equilibrium, the firm chooses to incentivize all workers to work and allocate the monitoring intensity evenly across workers at a given layer. This is because the firm can always reduce the cost by doing so, if these choices are not made. Lemma 1 proves and summarizes the above results.

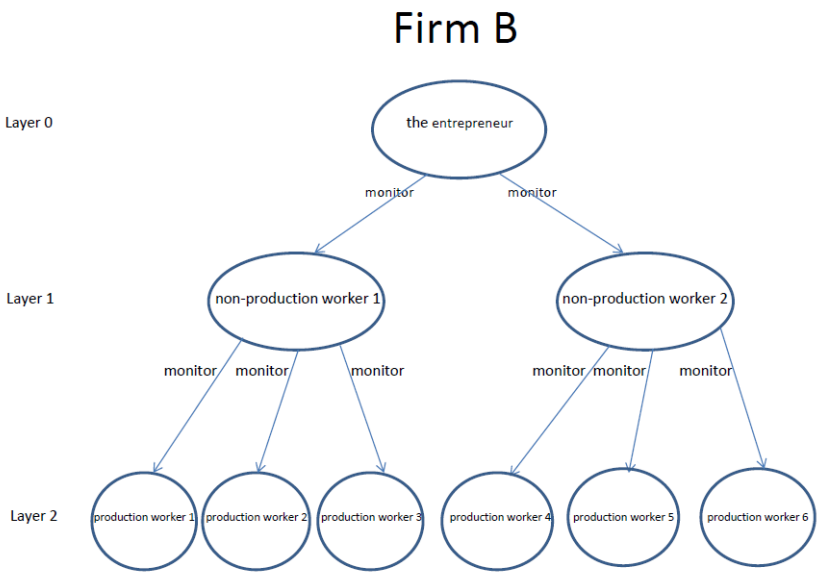
Lemma 1 *The firm incentivizes all workers to work (i.e., $a_i = 1$) and equalizes the monitoring intensity across workers at a given layer.*

Proof. See Appendix 7.2.1.

The entrepreneur faces the same incentive problem as her employees. She sits at the top of the hierarchy and is monitored by nobody. However, the entrepreneur of any firm that chooses to *stay* in the market is incentivized to exert effort in equilibrium (i.e., monitor her subordinates). First, staying in the market and shirking result in zero output

[17]Remember that there is a fixed number of entrepreneurs (i.e., one entrepreneur) at the top.

Figure 2: A Firm with Three Layers



and negative payoff for the entrepreneur. Second, the net payoff must be non-negative for the entrepreneur, if she decides to stay in the market and exert effort. This is because the entrepreneur would exit the market if the ex-post net payoff that is the difference between profit and the cost to exert effort were negative. Therefore, the entrepreneur exerts effort if she decides to stay in the market.

Now I characterize the firm's optimization problem. By substituting equation (7) into inequality (6), I derive the incentive-compatible wage for layer i as follows:

$$w_i(\theta) = \frac{\psi}{p_i(b, x_i(\theta))} = \psi b x_i. \tag{8}$$

The key feature of the above equation is that the incentive compatible wage $w_i(\theta)$ is negatively related to the supervision intensity. This relationship finds support in the data; see, for example, Rebitzer (1995) and Groshen and Krueger (1990).[18] Based on equations (1), (5) and (8) and Lemma 1, the optimization problem for a firm with the quality draw θ conditional on its staying in the market can be stated as

$$\begin{aligned} \max_{\{m_i\}_{i=1}^T, T} \quad & A\theta^{\frac{1}{\sigma}} m_T^{\frac{\sigma-1}{\sigma}} - \sum_{i=1}^{T} b\psi m_i x_i \\ s.t. \quad & x_i = \frac{m_i}{m_{i-1}}, \\ & m_0 = 1. \end{aligned} \tag{9}$$

[18]First, Rebitzer (1995) finds empirical evidence of a trade-off between supervision intensity and wage payment. Workers get paid less if they are under intensive monitoring. More importantly, Groshen and Krueger (1990) and Ewing and Payne (1999) find evidence on a negative relationship between the span of control and the wage paid to subordinates. This finding directly supports the basic trade-off of wage determination in the current model. Namely, a bigger span of control results in higher wage payment to subordinates.

where the first part of the above equation is the firm's revenue and the second part denotes the variable cost. The demand shifter A captures market size adjusted by the ideal price index and takes the following form:

$$A \equiv \Big(\frac{\gamma E}{P^{1-\sigma}}\Big)^{1/\sigma}, \tag{10}$$

where E is the total income of the economy. The number of entrepreneurs per firm is normalized to one, or, $m_0 = 1$. A big enough b is chosen to ensure that the probability of being monitored for any worker is always smaller than or equal to one.

The firm's optimal decisions given the number of layers can be solved in two steps. First, given an output level q, the first order conditions (FOCs) with respect to m_i's imply

$$w_T m_T = 2w_{T-1}m_{T-1} = ... = 2^{T-1}w_1 m_1, \tag{11}$$

where $m_0 = 1$ and $m_T = q$, as the number of production workers equals output q. This leads to the solution that

$$m_i(q,T) = 2^i\Big(\frac{q}{2^T}\Big)^{\frac{2^T-2^{T-i}}{2^T-1}}, \tag{12}$$

which is the number of workers at layer i. Thus, the firm's span of control at layer i is

$$x_i(q,T) = \frac{m_{i+1}(q,T)}{m_i(q,T)} = 2\Big(\frac{q}{2^T}\Big)^{\frac{2^{T-(i+1)}}{2^T-1}}, \tag{13}$$

which increases with q given the number of layers. Equation (12) shows that employment at each layer increases with output given the number of layers. Moreover, equation (13) indicates that the number of workers increases disproportionately more at upper layers, which leads to a bigger span of control at each layer. This is due to the fixed number of entrepreneurs at the top.

Second, optimizing over output yields

$$q(\theta,T) = m_T(\theta,T) = \left[\frac{A\beta\theta^{\frac{1}{\sigma}}}{b\psi 2^{2-\frac{T}{2^T-1}}}\right]^{\frac{\sigma(2^T-1)}{\sigma+(2^T-1)}}, \tag{14}$$

which is the firm's optimal output level as well as the number of production workers. Substituting equations (12) and (14) into equation (9) leads to the firm's operating profit (i.e., profit before paying the fixed cost) and revenue as

$$\pi(\theta,T) = (1-\frac{\beta(2^T-1)}{2^T})(A\theta^{\frac{1}{\sigma}})^{\frac{2^T\sigma}{\sigma+(2^T-1)}}\left(\frac{\beta/b\psi}{\Big(2^{\frac{2^{T+1}-2-T}{2^T-1}}\Big)}\right)^{\frac{(\sigma-1)(2^T-1)}{\sigma+(2^T-1)}} \tag{15}$$

and

$$S(\theta,T) = (A\theta^{\frac{1}{\sigma}})^{\frac{2^T\sigma}{\sigma+(2^T-1)}}\left(\frac{\beta/b\psi}{\Big(2^{\frac{2^{T+1}-2-T}{2^T-1}}\Big)}\right)^{\frac{(\sigma-1)(2^T-1)}{\sigma+(2^T-1)}}, \tag{16}$$

which will be used later. the firm's employment, output, and revenue increase continuously with the quality draw θ given the number of layers. More importantly, all of these variables increase *discontinuously* when the firm adds a layer as shown below. With the firm's optimal decisions on employment and output in hand, I can solve for the optimal number of layers, which is the final step of the firm's optimization problem.

2.3 Endogenous Selection into the Hierarchy with Different Numbers of Layers

This subsection characterizes a firm's cost functions in order to solve for the optimal number of layers in a firm's hierarchy. The key result is that firms with better quality draws choose to have more layers and produce more in equilibrium.

Consider a firm that produces q units of output. The variable cost function of such a firm is given by[19]

$$TVC(q,b) = \min_{T \geq 1} TVC_T(q,b),$$

where $TVC(q,b)$ is the minimum variable cost of producing q and $TVC_T(q,b)$ is the minimum variable cost of producing q using the hierarchy with $T+1$ layers. Based on equations (12) and (13), the minimum cost $C_T(q,b)$ can be derived as

$$TVC_T(q,b) = \sum_{i=1}^{T} m_i(q,T) w_i(q,T) = \sum_{i=1}^{T} b\psi \frac{m_i^2(q,T)}{m_{i-1}(q,T)} = (2 - \frac{1}{2^{T-1}}) b\psi 2^{1-\frac{T}{2^T-1}} q^{\frac{2^T}{2^T-1}}. \tag{17}$$

Therefore variable costs given the number of layers increase with output. More importantly, better MT, which is denoted by a smaller value of b, pushes down the variable cost of production given any number of layers. This is because the firm can incentivize workers to exert effort by paying lower costs.

With the firm's cost functions given different numbers of layers in hand, I can characterize properties of the AVC curve and the MC curve using the following proposition.

Proposition 1 *Given the number of layers, both the average variable cost and the marginal cost increase continuously with output. The average variable cost curve kinks and its slope decreases* discontinuously *at the output level where the firm adds a layer. As a result, firms that produce more have more layers. The marginal cost falls discontinuously when the firm adds a layer.*

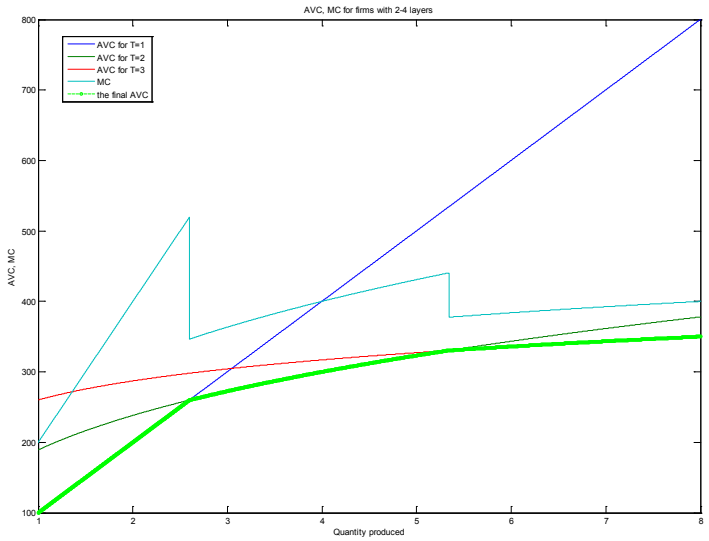
Proof. See Appendix 7.2.2.

Figure 3 illustrates the AVC curve and the MC curve. The AVC curve denoted by the bold green curve is the lower envelope of all the AVC curves given different numbers of layers. The MC curve does not increase with output monotonically. The span of control increases at all layers when the firm increases output and keeps the number of layers unchanged. Therefore, wages increase at all layers, which implies that both the AVC and the MC increase with output given the number of layers. Wages fall at existing layers when the firm adds a layer owing to the smaller span of control. This leads to a discontinuous decrease in the MC. Because of this drop, the AVC curve kinks and its slope decreases *discontinuously* at the output level where the firm adds a layer.

Proposition 1 establishes a positive relationship between output and the optimal number of layers. When the output level is low, it is ideal to have a smaller number of layers.

[19]For firms that have one layer (i.e., $T = 0$ or self-employed entrepreneurs), the hierarchy and MT are not needed. As this paper focuses on the hierarchy and MT, I do not consider these firms in what follows.

Figure 3: Average Variable Cost and Marginal Cost



This is because adding a layer is like an investment that reduces the MC at the expense of a fixed cost. This property of the AVC curve is evident in Figure 3, as the AVC curve given a bigger number of layers has a smaller slope and a larger intercept on the y axis. Similarly, it is optimal to have more layers when output is high. In summary, the number of layers and output increase hand in hand in equilibrium.

What is the relationship between the firm's demand draw and the optimal number of layers? The key observation is that it is more profitable for a firm with a better demand draw to add a layer. This is due to the key feature of the AVC curve that adding a layer is like investing a fixed amount of money to reduce the MC. In other words, there is a complementarity between the level of the firm's demand draw and its incentive to add a layer. Proposition 2 characterizes a positive relationship between the firm's demand draw and the optimal number of layers by proving this complementarity.

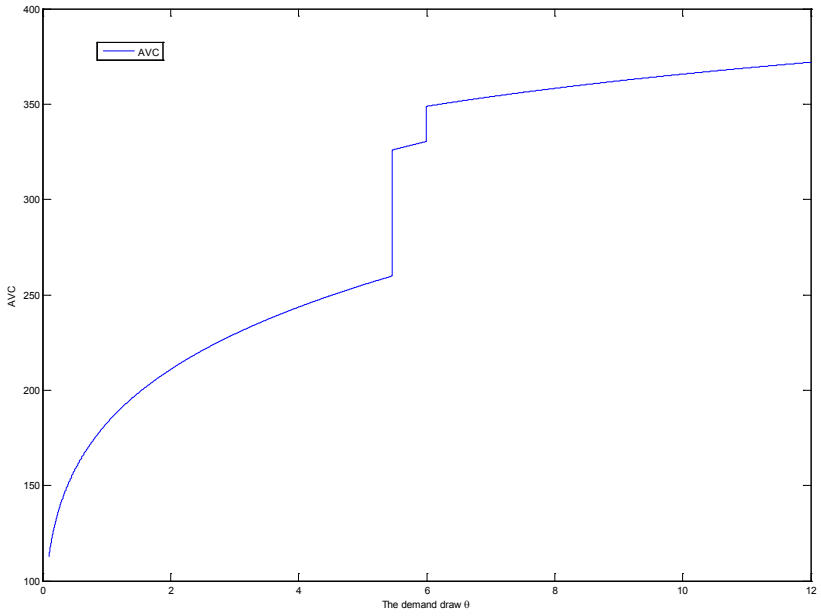
Proposition 2 *Firms that receive better demand draws have more layers.*

Proof. See Appendix 7.2.3.

Proposition 2 and the distinctive feature of the AVC curve discussed above are the keys to understanding the pro-competitive effect of an improvement in MT. Improved MT reduces labor cost and incentivizes firms to grow. Furthermore, it incentivizes firms with better demand draws to expand more and benefits them disproportionately more. This is because firms with better demand draws have more layers, and the elasticity of the AVC with respect to output is smaller for these firms.

The theoretical results proven in Proposition 2 are consistent with the empirical findings from Caliendo, Monte, and Rossi-Hansberg (2012). First, firms that are bigger in terms of either employment or value added are found to have more layers in the data set of French firms. Second, firms that grow are found to (weakly) increase the number of

Figure 4: Average Variable Cost and the Demand Draw



layers as well. All this evidence supports a key result of this paper: firms with better demand draws have more layers.

I close this subsection by discussing how price and firm size respond to the change in the firm's demand draw. These results are useful, as I will analyze the firm size distribution in the next section. Proposition 3 summarizes the results.

Proposition 3 *Given the number of layers, output, employment, and price increase continuously with the firm's demand draw. When the firm adds a layer, output and employment increase* discontinuously*, while price falls* discontinuously*.*

Proof. See Appendix 7.2.4.

The strategy of a firm to grow depends crucially on whether or not the production is reorganized. When the firm grows as a result of an improvement in the quality of its product and keeps the number of layers unchanged, price increases as the MC increases. However, when one layer is added, the MC falls, which leads to lower prices. Exactly because of this discontinuous decrease in the MC, firm size increases discontinuously when production is reorganized. As a result, the AVC as a function of the demand draw jumps when the firm adds a layer, as shown in Figure 4.[20]

[20]The optimal output level is substituted into the firm's AVC for calculating the AVC as a function of the demand draw.

2.4 The Spans of Control, Wages, and Relative Wages

The incentive-based hierarchy proposed above has predictions for firm-level outcomes. This subsection presents predictions on the spans of control, wages, and relative wages. The first two variables increase with the demand draw given the number of layers, and they decrease discontinuously when firms add a layer as in Caliendo and Rossi-Hansberg (2012). In addition, relative wage, defined as the ratio of the supervisor's wage to his direct subordinate's wage, behaves in the way consistent with the findings of Caliendo, Monte, and Rossi-Hansberg (2012).

What happens to the firm-level outcomes when the firm expands due to an improvement in the quality of its product and keeps the number of layers unchanged? Proposition 4 summarizes the results.

Proposition 4 *Given the number of layers, both the span of control and wages increase with the firm's quality draw at all layers. Furthermore, relative wages increase with the firm's quality draw at all layers as well.*

Proof. See Appendix 7.2.5.

The change in the span of control is the key to understanding this proposition. When the firm is constrained to keep the number of layers unchanged, the only way to expand is to increase the span of control at all layers. When the span of control is larger, monitoring is less effective, which implies that higher wages are needed to incentivize workers. Furthermore, wages increase disproportionately more at upper layers. The share of workers at upper layers in total employment decreases when the firm grows without adjusting the number of layers. Thus, the firm tolerates disproportionately more increases in wages at upper layers while keeping increases in wages at lower layers relatively small.

The results of Proposition 4 are consistent with the evidence presented in Caliendo, Monte, and Rossi-Hansberg (2012). First, given the number of layers, wages are found to increase with firm size in both cross-sectional and time-series regressions. These results are what the model predicts, as the bigger demand draw leads to bigger firm size, as Proposition 3 shows. Second, and more importantly, relative wages are shown to increase with firm size at all layers when the firm does not change the number of layers. This prediction implies that, although workers all gain when a firm expands without reorganization, workers at higher layers gain more. This unambiguous prediction is a unique prediction of my model, as the model presented in Caliendo and Rossi-Hansberg (2012) is silent on how relative wages change when the firm expands.

When the firm chooses to add a layer owing to an improvement in the quality of its product, the firm-level outcomes move in the opposite direction, as summarized by the following proposition.[21]

Proposition 5 *When the firm adds a layer owing to a marginal improvement in the quality of its product, both the span of control and wages fall at existing layers. Furthermore, relative wages decrease at existing layers as well.*

[21] Following Caliendo and Rossi-Hansberg (2012), I assume that the firm adds a layer from above. As the entrepreneur is at layer zero, layer i becomes layer $i+1$ where $i \geq 1$, when the entrepreneur adds a layer.

Proof. See Appendix 7.2.6.

The change in the span of control is again the key to understanding this proposition. When the firm expands by adding a layer, the constraint at the top (i.e., the fixed supply of entrepreneurs) is relaxed. Thus, the firm can expand and economize on its labor cost at the same time. As a result, the span of control decreases at existing layers, which leads to lower wages paid to employees at existing layers. On top of that, wages fall disproportionately more at upper layers. The share of workers at upper layers in total employment increases because of the shrinking span of control. Consequently, it is an efficient way to economize on labor cost by reducing their wages disproportionately more.

The results of Proposition 5 are largely consistent with the evidence presented in Caliendo, Monte, and Rossi-Hansberg (2012) as well. First, wages are found to decrease at existing layers when firms expand by adding a layer. Second, relative wages fall at existing layers as well for firms that expand and add a layer. In total, the model's predictions on wages and relative wages are consistent with the empirical findings presented in Caliendo, Monte, and Rossi-Hansberg (2012).

2.5 Firm Productivity

I discuss firm productivity and its relationship with output given the quality of MT here. Following Caliendo and Rossi-Hansberg (2012), I use the inverse of unit costs (i.e., output divided total costs) to measure firm productivity. As firm productivity is simply the inverse of unit costs, I analyze how the unit costs vary with output as well as the number of layers first.

Formally, the unit costs given a number of layers and the inefficiency of MT are defined as

$$UC_T(q,b) \equiv \frac{TVC_T(q,b) + f_0}{q} = AVC_T(q,b) + AFC(q), \tag{18}$$

where f_0 is the fixed production cost and $AFC(q)$ is the average fixed cost. $UC_T(q,b)$ is a function of output given T and b. Second, the unit costs given the inefficiency of MT are defined as

$$UC(q,b) \equiv \frac{TVC(q,b) + f_0}{q} = AVC(q,b) + AFC(q). \tag{19}$$

As the fixed production cost does not affect the firm's choice for the number of layers, equation (19) can be restated as

$$UC(q,b) = UC_T(q,b), \quad \forall q \in [q_{T-1}, q_T),$$

where q_T is defined as the solution to $AVC_T(q_T,b) = AVC_{T+1}(q_T,b)$. In what follows, I discuss how $UC_T(q,b)$ and $UC(q,b)$ vary with output.

First, the curve of unit costs given the number of layers and the inefficiency of MT has an "U" shape. In other words, it decreases first and increases afterwards. Note that the average fixed cost (AFC) always decreases with output, while the AVC always increases with output. The decrease in AFC dominates the increase in AVC when output increases from a low level and vice versa. Thus, the unit costs given a number of layers decrease until output exceeds a certain level. Furthermore, the slope of the curve approaches

zero when output goes to infinity, as both the decrease in AFC and the increase in AVC triggered by an increase in output become infinitesimally small.

Second, I discuss the relationship between the curves of unit costs given various numbers of layers. I define the minimum efficient scale (MES) given a number of layers as the scale of production at which a firm minimizes unite costs given a number of layers, and the minimum unit costs (MUC) given a number of layers as the unit costs when the scale of production is at its MES. Mathematically, the MES given b and T is defined as

$$q_{Tm}(b) \equiv argmin_q \, UC_T(q, b). \tag{20}$$

And the MUC given b and T is defined as

$$MUC_T(b) \equiv UC_T(q_{Tm}(b), b). \tag{21}$$

What is the relationship between various $MUC_T(b)$ given different numbers of layers? The following assumption assures that $MUC_T(b)$ decreases with T, which implies that the hierarchy with more layers has the lower MUC.

Assumption 1 $f_0 > 4b\psi$.

I adopt the above assumption in this paper. Under this assumption, firm productivity has an increasing overall trend with respect to output, as it is the inverse of the unit costs. On the contrary, firm productivity has an decreasing overall trend with respect to output, if Assumption 1 is violated.[22] Essentially, Assumption 1 requires that MT is efficient enough (i.e., b is small enough).

Now, I characterize the properties of $q_{Tm}(b)$, $MUC_T(b)$, and $UC_T(q, b)$ using the following proposition.

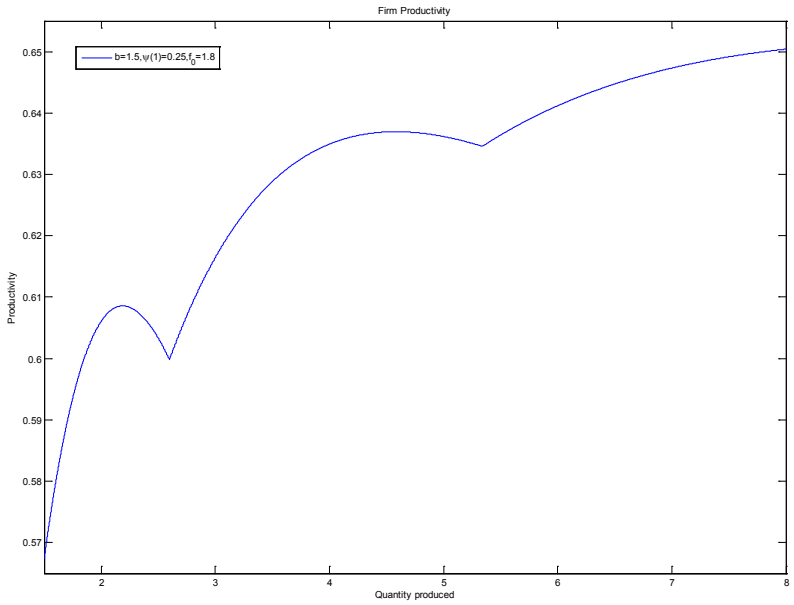
Proposition 6 *Given the number of layers and the inefficiency of MT, the curve of unit costs is "U"-shaped, and the slope of it approaches zero when output goes to infinity. Under Assumption 1, the MES given the number of layers increases with the number of layers; the MUC given the number of layers decreases with the number of layers.*

Proof. See Appendix 7.2.7.

The inefficiency of MT is the key to understanding this proposition. improved MT makes the MES increase and the MUC decrease given a number of layers. Moreover, firms with more layers gain disproportionately more from such an improvement, as the share of the total variable cost in total costs is bigger for these firms. Thus, the MES increases more and the MUC decreases more for firms with more layers after an improvement in

[22] One implication here is that the positive correlation between firm productivity and size is stronger in economies with better MT, and this correlation may be negative when the quality of MT is sufficiently low. Interestingly, Bartelsman, Haltiwanger, and Scarpetta (2013) find that the positive correlation between firm productivity and size is much stronger in developed countries such as the U.S. and Germany compared to developing countries such as Hungary and Slovenia. In Romania, covariance between firm productivity and size is even negative. Although there are no management data of firms in Hungary, Slovenia and Romania available, firms in Poland that is also a central European country have worse MT compared to firms in the U.S. and Germany. Therefore, the finding in Bartelsman, Haltiwanger, and Scarpetta (2013) can be seen as a supportive evidence for my productivity measure.

Figure 5: Firm Size and Productivity



MT. As a result, the MES increases with the number of layers, and the MUC decreases with the number of layers when MT is efficient enough.

As firm productivity is simply the inverse of its unit costs, I characterize the overall shape of the curve of firm productivity using the following corollary.

Corollary 1 *The curve of firm productivity is a combination of various inversely "U"-shaped curves. More importantly, the overall trend is that firm productivity increases with output.*

Figure 5 illustrates how firm productivity varies with output. It is evident that the overall trend is that firm productivity increases with output. This implies that small firms are less productive on average. Aggregate productivity of firms increases if the smallest firms are driven out of the market after some shock. This is one effect of improved MT on aggregate productivity that I will analyze in the next section.

2.6 General Equilibrium

I close the model by aggregating across firms and solve for the general equilibrium in this subsection. There are two product markets and one labor market. Entrepreneurs decide whether or not to enter the CES sector and must be indifferent between entering or not in equilibrium due to the large pool of potential entrepreneurs. Workers choose which sector and which labor submarket to enter.[23] They must be indifferent between different labor markets (and submarkets) in equilibrium, as they can freely move across sectors and firms.

[23] I will explain what labor submarkets mean in what follows.

2.6.1 Product Market Equilibrium

There are two equilibrium conditions for the CES sector: the zero cutoff *payoff* (ZCP) condition and the FE condition. They are used to pin down two equilibrium variables: the exit cutoff for the quality draw (i.e., $\bar{\theta}$) and the mass of firms in equilibrium (i.e., M). First, the ZCP condition that firms with the quality draw $\bar{\theta}$ earn zero payoff can be written as

$$\Pi(\bar{\theta}, A) = 0, \tag{22}$$

where $\Pi(\bar{\theta}, A) \equiv \pi(\bar{\theta}, T(\bar{\theta}, A), A) - (f_0 + \psi)$ is the entrepreneur's payoff. This condition pins down the exit cutoff $\bar{\theta}$ given the adjusted market size A. Note that the ZCP condition here incorporates both the fixed cost to produce and the cost of exerting effort, as entrepreneurs of active firms exert effort to monitor their subordinates in equilibrium. For simplicity, I use $f \equiv f_0 + \psi$ to denote the overall "fixed cost" to produce.

The FE entry condition implies that the expected payoff from entering the CES sector equals the outside option of entrepreneurs, or

$$\int_{\bar{\theta}}^{\infty} \Pi(\theta, A) g(\theta) d\theta = f_e, \tag{23}$$

where $f_e \equiv f_1 + f_2$ is the overall opportunity cost to enter the CES sector, and $g(\theta)$ is the probability density function (PDF) of the quality draw θ. This equation determines the adjusted market size A given the exit cutoff $\bar{\theta}$.

The mass of firms is undetermined in the homogeneous sector, and the managerial talent is not needed for firms in the homogeneous sector. Given the assumptions of a linear production technology and perfect competition in the homogeneous sector, firm boundaries are not defined in that sector. Therefore, I assume that entrepreneurs choose whether or not to enter the CES sector. In equilibrium, the FE condition holds with equality if and only if

$$N \geq \frac{M}{1 - G(\bar{\theta})},$$

where M is the mass of active firms in equilibrium, and $G(\theta)$ is the cumulative distribution function (CDF) of θ. A sufficiently large N ensures that the above inequality holds.

The equilibrium condition for the homogeneous sector is that supply of the homogeneous good equals the demand for it, or

$$p_h L_h = (1 - \gamma) E, \tag{24}$$

where L_h is the number of workers in the homogeneous sector. This condition pins down p_h, which is the price of the homogeneous good as well as workers' wages in this sector.

2.6.2 Labor Market Equilibrium

The labor market in the CES sector is characterized by competitive search. Firms demand workers for each layer, and a worker chooses one type of job to apply for in order to maximize the expected payoff. Firms randomly select workers among those who come to apply for jobs to employ. A type of job corresponds to a firm-layer pair (θ, i), as different firms offer different wages for various positions (i.e., layers). In other words, there are

labor submarkets indexed by (θ, i) in the CES sector. As workers are homogeneous and can freely choose which type of job to apply for, the expected payoff from applying for any type of job must be the same in equilibrium. Moreover, this uniform expected payoff must be equal to the wage offered in the traditional sector, which is the outside option of workers entering the CES sector. In total, I have

$$\frac{m_i(\theta)}{Q(\theta,i)}(w_i(\theta)-\psi)=\frac{m_{i'}(\theta')}{Q(\theta',i')}(w_{i'}(\theta')-\psi)=p_h \quad \forall\,(i,i')\,\forall\,(\theta,\theta'), \tag{25}$$

where $m_i(\theta)$ is the firm's labor demand at layer $i(\geq 1)$, and $Q(\theta,i)$ is the number of workers who come to apply for this type of job. $\frac{m_i(\theta)}{Q(\theta,i)}$ is the probability of being employed in labor submarket (θ,i), and $(w_i(\theta)-\psi)$ is the net payoff of being employed. Different job turn-down rates across labor submarkets (i.e., $\frac{Q(\theta,i)-m_i(\theta)}{Q(\theta,i)}\geq 0$) are needed to equalize the expected payoff from entering various labor submarkets. As a result, there is unemployment in equilibrium.

I derive the labor-market-clearing condition in two steps. First, the number of workers who choose to enter the CES sector (i.e., L_c) can be derived from the workers' indifference condition in equation (25), or

$$\begin{aligned} L_c &= \int_{\theta=\bar{\theta}}^{\infty}\sum_{i=1}^{T(\theta,A)} Q(\theta,i)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta \\ &= \frac{WP(\bar{\theta},A,M)-\psi LD(\bar{\theta},A,M)}{p_h}, \end{aligned} \tag{26}$$

where $WP(\bar{\theta},A,M)$ is the total wage payment in the CES sector, and $LD(\bar{\theta},A,M)$ is the number of workers employed in the CES sector,[24] or

$$LD=\int_{\theta=\bar{\theta}}^{\infty}\sum_{i=1}^{T(\theta,A)} m(\theta,i)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta. \tag{27}$$

Equation (26) has a intuitive explanation. It says that the *total* expected payoff of workers entering the CES sector (i.e., p_hL_c due to the indifference condition) is equal to the difference between the total wage payment and the total disutility to exert effort.

Second, the labor-market-clearing condition indicates that the number of workers employed in the homogeneous sector is the difference between the endowment of labor and the number of workers who choose to enter the CES sector, or

$$L_h=L-L_c. \tag{28}$$

Equations (26) and (28) are two labor market equilibrium conditions that are used to determine the allocation of labor between two sectors.

2.6.3 Equilibrium and Unemployment

The market-clearing condition of the final composite good implies that

$$E=\int_{\bar{\theta}}^{\infty} LC(\theta)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta+p_hL_h+\Big[f_0+f_1\big(\frac{\bar{\theta}}{\theta_{min}}\big)^k\Big]M+\Big[\psi+f_2\big(\frac{\bar{\theta}}{\theta_{min}}\big)^k\Big]M, \tag{29}$$

[24] For further discussion of the labor market equilibrium in the CES sector, see Appendix 7.2.8.

where $LC(\theta)$ is the total labor cost (i.e., total wage payment) of firms with the demand draw θ. The third part of the right hand side (RHS) of equation (29) is the demand for the final composite good by firms, and the last part of the RHS of equation (29) is the consumption of active entrepreneurs who earn profit in equilibrium.[25] The aggregate income of the economy equals the aggregate expenditure which includes two part: the demand from workers and from firms. Note that only workers and *active* entrepreneurs demand goods to consume. Entrepreneurs who choose not to enter the CES sector enjoy their outside option without consuming goods, and entrepreneurs who enter the CES sector and choose not to produce don't consume goods as their income is zero.

The general equilibrium of this economy is characterized by the quality threshold of the firm that obtains zero payoff, $\bar{\theta}$, the mass of firms that operate M, the price of the homogeneous good p_h, the labor allocation between two sector, L_c and L_h, and the aggregate income E. These six equilibrium variables are obtained by solving the six equations (i.e., equations (22), (23), (24), (26), (28) and (29)). Obviously, one equilibrium condition is redundant due to Walras' law, and I normalize the price of the final composite good in equation (5) to one.

One implicit assumption for the existence of the equilibrium is that the probability of being employed implied by equation (25) is smaller than or equal to one in *every* labor submarket in equilibrium (i.e., $\frac{m_i(\theta)}{Q(\theta,i)} \leq 1 \ \ \forall \ (i,\theta)$). In other words, wages offered in the CES sector must satisfy

$$w_i(\theta) - \psi \geq p_h \ \ \forall \ (i,\theta), \tag{30}$$

where $w_i(\theta)$ is determined in equation (8). The above inequality would be violated if ψ were zero. Firms do not need to hire non-production workers and pay incentive-compatible wages to them, if exerting effort does not generate any cost to the workers. At the same time, there is no unemployment in *all* labor submarkets, and every worker in the CES sector receives the same wage if ψ were zero. I don't consider this case in the paper, as MT does not matter in this case. In the paper, I focus on the case in which unemployment exists in *every* labor submarket, and the incentive-compatible wage determined in equation (8) satisfies the constraint specified in equation (30) in *every* labor submarket. There are three reasons why I want to investigate this type of equilibrium. First, the model delivers clean insights and testable implications in this case. Second, the testable implications at the micro-level (e.g., wages, relative wages, and the choice of the number of layers etc.) have shown to be consistent with the evidence presented in Caliendo, Monte, and Rossi-Hansberg (2012). Finally, as what we are going to show, the model's predictions at the aggregate level (i.e., the effect of the quality of MT on the firm size distribution and decentralization decisions of firms) are consistent with the evidence presented in Hsieh and Klenow (2009, 2012) and Bloom et al. (2013) as well. The following proposition discusses the existence and uniqueness of an equilibrium with unemployment in every labor submarket.[26]

[25]The equilibrium condition stated in equation (29) has used the FE condition described above. Ex post profit per active firm must compensate both the cost to exert effort and the forgone outside option. The overall forgone utility is $f_2\big(\frac{\bar{\theta}}{\theta_{min}}\big)^k M$, and total utility cost to exert effort is ψM. Therefore, profit per active firm which compensates these costs is $\psi + f_2\big(\frac{\bar{\theta}}{\theta_{min}}\big)^k$.

[26]When the outside option of workers (i.e., p_h) is not too small in equilibrium, the equilibrium has the property that some labor submarkets have unemployment and the others don't. (i.e., constraint (30) is violated for some labor submarkets). In this case, firms that would offer lower incentive-compatible

Proposition 7 *When* $\frac{\sigma-1}{\sigma} \neq \gamma$*, there is a range of parameter values in which a unique equilibrium with unemployment in every labor submarket exists.*

Proof. See Appendix 7.2.8.

Admittedly, the condition assuring the existence of a unique equilibrium with unemployment in every labor submarket involves endogenous variables.[27] This is because both wages offered in the CES sector and wage offered in the homogeneous sector cannot be solved analytically. However, when I treat the number of layers as a continuous variable as in Keren and Levhari (1979) and Qian (1994), the condition can be stated using exogenous parameters only. Readers are referred to the online appendix for more details.

When $\frac{\sigma-1}{\sigma} = \gamma$, the aggregate labor demand that takes into account the product-market-clearing conditions does not respond to the change in the expected payoff of workers (i.e., p_h). Thus, there is either no equilibrium or infinitely many equilibria depending on values of other parameters in this knife edge case. Thanks to the uniqueness of the equilibrium under restrictions on parameter values, I can analyze how an improvement in MT affects various economic activities.

I close this section by discussing the role of unemployment rate in this model. Although the wage determination in the current model is similar to the one used in the efficiency wage theory (e.g., Shapiro and Stiglitz (1984)), the role of unemployment rate is different. In the efficiency wage theory, the aggregate unemployment *rate* feeds back to the incentive-compatible wage in a dynamic setup, and unemployment is present only when there exists an exogenous separation rate between firms and workers.[28] In this paper, unemployment (or being fired) still serves as a disciplinary device to incentivize workers to exert effort. However, unemployment *rates* (more precisely, job-turn-down rates) across different labor submarkets are used to equalize the expected payoffs from entering labor submarkets in the CES sector and the homogeneous sector. Essentially, the role of unemployment rate in my model is the same as in Harris and Todaro (1970) and is similar to the role of the labor market tightness in the literature on competitive search (e.g., Moen (1997) etc.).[29]

---

wages in the absence of the outside option of workers are forced to raise wages up to $p_h + \psi$. I discuss this case in the online appendix and show that qualitative results of the model in this case are the same as the ones we are going to derive in the paper.

[27] The condition assuring the existence and uniqueness of the equilibrium is that $\frac{\sigma-1}{\sigma} \neq \gamma$ and $p_h \leq w_{min}(\bar{\theta}) - \psi$, where $w_{min}(\bar{\theta})$ is the lowest wage among wages offered in the CES sector. This is a sufficient and necessary condition for the existence and uniqueness of the equilibrium.

[28] Remember that in Shapiro and Stiglitz (1984) every agent is incentivized to work.

[29] More specifically, the difference in the role of unemployment rate between the efficiency wage theory and my model comes from the assumption of how the firm punishes misbehaving workers. In my static model, the firm punishes shirking workers whose misbehavior has been detected by firing them and reducing the wage payments. In the efficiency wage theory, the firm punishes shirking workers whose misbehavior has been detected by firing them but *not* reducing the wage payments. Thus, the incentive to work comes from a decrease in the future wage income due to unemployment in the efficiency wage theory. It is probably true that workers do get wage cuts when their performance does not meet some goals that have been preset in practice.

3 Management Technology, Institutional Quality, and Firm-Level Outcomes

This section investigates how an improvement in MT affects firm characteristics as well as welfare. Ample evidence suggests that there are substantial differences in the quality of MT across countries due to factors that are beyond the control of firms. Furthermore, Hsieh and Klenow (2009) show that China and India whose firms receive low management scores have more firms of a small size and are worse at getting efficient firms to obtain big market shares compared with the U.S. Finally, the internal structure of firms differs across countries due to differences in MT and affects firm size and performance as well. Namely, firms in India are, compared with those in the U.S., less decentralized owing to worse MT and weak enforcement of laws, and the low level of decentralization impedes Indian firm expansion (e.g., Bloom, Sadun, and Van Reenen (2012 a), Bloom et al. (2013)). The purpose of this section is to show that there is a link between the quality of MT and the firm size distribution as well as the internal structure of firms.

This section is divided into two subsections. In the first subsection, I consider the case in which the quality of MT only differs across countries and investigates how improvements in management quality affect economics outcomes. In the second subsection, I consider the case in which management quality differs both across firms and countries and derive a testable hypothesis that will be tested in Section 5.

3.1 Pro-Competitive Effect of Better Management Technology

I consider a scenario in which MT that is common across all firms improves. Such an improvement is equivalent to a decrease in b in the model, as it becomes easier for the firm to catch and fire shirking workers after the change. As a result, firms' labor costs decrease, since workers' wages are determined by the incentive compatibility constraint.

An improvement in MT generates a pro-competitive effect that reallocates resources toward more efficient firms. This improvement favors more efficient firms, as they have more layers. More specifically, an improvement in MT benefits all firms as it reduces firms' labor cost. Moreover, firms with more layers gain disproportionately more, as their AVCs increase less rapidly with output. More precisely, the AVC functions of firms with more layers have *smaller* elasticities with respect to output due to the drop in the MC when firms add a layer. As a result, firms with the worst demand draws are forced to leave the market; firms whose demand draws are in the middle range receive shrinking revenue and profit; and firms with the best demand draws expand. In other words, improvements in MT facilitate inter-firm resource allocation by favoring bigger firms, which is what Bloom et al. (2013) argue in their paper.

Endogenous selection of a hierarchy with a specific number of layers is the key to understanding the pro-competitive effect of the better management technology. In a hypothetic world, if all firms were forced to have the same number of layers, the uneven effect across firms would disappear, as all firms would have the same AVC function. As a result, the exit cutoff for the quality draw θ is unaffected by an improvement in MT. Furthermore, firms' revenue and profit are unchanged as well. In summary, the pro-competitive effect of improved MT is present only when firms *endogenously* choose to have different numbers of layers.

How does the internal organization of firms evolve when MT improves? Bloom et al. (2013) find that Indian firms are unwilling to decentralize their production processes (i.e., constrained span of control), as it is hard to catch and punish misbehaving employees in India. Furthermore, they argue that poor monitoring and weak enforcement of laws are reasons for why Indian firms can't catch and punish misbehaving workers easily. Finally, they argue that low level of decentralization is one reason for why Indian firms are small on average. My model gives an explanation for these findings. First, when firms are able to monitor their employees more easily, the span of control increases given the number of layers. Second, and more importantly, the number of layers also increases weakly for each firm because of better monitoring. Firms expand when monitoring becomes more effective, and the expansion incentivizes firms to have more layers.[30] As a result, firms have fewer layers or constrained span of control in economies with worse MT. Furthermore, less decentralized production processes are associated with smaller average firm size as what I will show next. In total, firms are less decentralized in economies with poor MT, and low decentralization is one reason for why firms are small in these economies.

In order to derive analytical results on the firm size distribution and the distribution of the number of layers, I assume that θ follows a Pareto distribution with a coefficient k, or

$$G(\theta) = 1 - \left(\frac{\theta_{min}}{\theta}\right)^k. \tag{31}$$

The following proposition summarizes the changes in firm characteristics due to an improvement in MT. Note that the above distributional assumption is only needed for the results on the firm size distribution and the distribution of the number of layers in Proposition 8.

Proposition 8 *Suppose management technology that is common across all firms improves. Consider the case in which the minimum number of layers among active firms is unchanged. For the economy as a whole, the exit cutoff for the quality draw increases. At the firm level, all surviving firms either increase the number of layers (weakly) or make the span of control bigger and keep the number of layers unchanged. Finally, if the quality draw follows a Pareto distribution, both the firm size distribution and the distribution of the number of layers move to the right in the FOSD sense.*

Proof: See Appendix 7.2.9.

I focus on the case in which the minimum number of layers of active firms is unchanged when MT improves, although similar results emerge in the other cases. The reason why I focus on this case is that there are always some extremely small firms that have only two layers (i.e., $T = 1$) in every economy of the world. Therefore, the case considered in the paper is empirically more relevant. Furthermore, I prove that all the results of Proposition 8 except for the prediction on the span of control hold, when the number of layers is treated as a continuous variable as in Keren and Levhari (1979) and Qian

[30]Note that output and employment go up for *all* firms, but revenue and operating profit fall for the smallest firms after MT improves.

(1994).[31] In total, the aggregate-level predictions (i.e., changes in the exit cutoff, the firm size distribution, and the distribution of the number of layers) in the case of the continuous number of layers are qualitatively the same as the aggregate-level predictions in the case of the discrete number of layers.

The FOSD results have important implications for resource allocation in the economy and are consistent with the data. First, the FOSD result for the firm size distribution means that there are fewer small firms in economies with superior MT. Furthermore, firms with better demand draws have bigger sales, and the average firm size is bigger in such economies as well. These theoretical predictions show the key role of an improvement in MT. That is resources are reallocated toward more efficient firms. Second, the FOSD result for the distribution of the number of layers implies that firms are less decentralized in economies with worse MT, which is one important reason for why firms with the best demand draws don't have big enough sales in such economies. Finally, these theoretical predictions are consistent with several key findings in Hsieh and Klenow (2009, 2012) and Bloom et al. (2013) that are discussed in the introduction.[32]

Other than changes in the firm size distribution and the internal organization of firms, the weighted average of firm productivity also increases as a result of an improvement in MT. Gains in the weighted average of firm productivity come from three sources. First, firms with the worst demand draws which are less productive exit the market after an improvement in MT (i.e., the between-firm effect). Second, market shares of more productive firms increase, as improved MT favors more productive firms. This makes the weighted average of firm productivity increase as well (i.e., the between-firm effect). Finally, productivity of all surviving firms increases, as improved MT reduces firms' costs (i.e., the within-firm effect). The above three sources of gains in the weighted average of firm productivity are shown to be present as well, when the number of layers is treated as a continuous variable.[33] In total, these three effects together increase the weighted average of firm productivity, as shown in Figure 6.

Other than firm-level outcomes, I am also interested in how improved MT affects the workers' welfare. Since entrepreneurs' outside option is exogenous, I focus on how improvements in MT affect welfare of workers. As workers can freely move between two sectors, the expected payoff from entering the CES sector must equal the wage offered in the homogeneous sector. Therefore, the (ex ante) expected payoff of workers entering the CES sector is a sufficient statistic to evaluate welfare. In what follows, I discuss how this changes when MT improves.

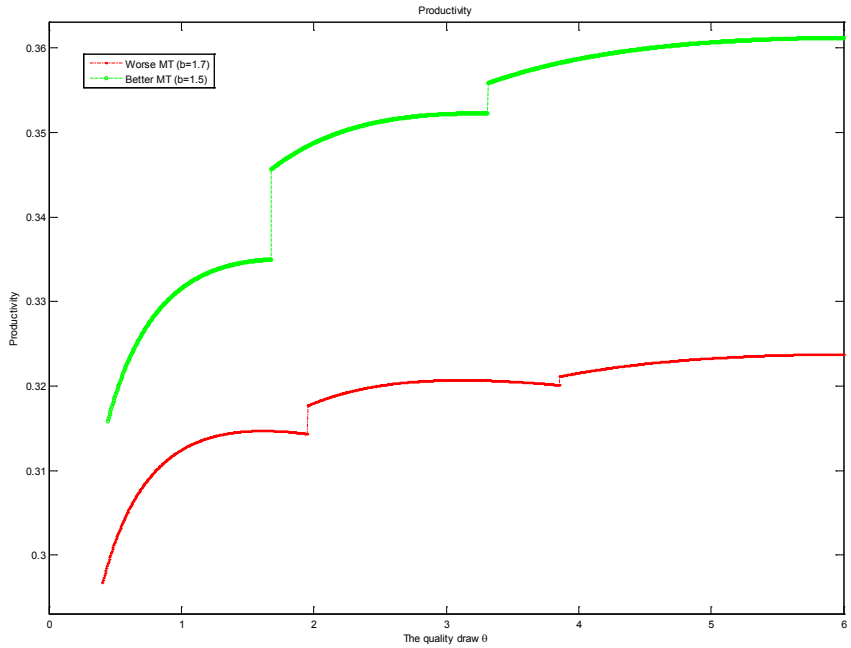
Better MT can either increase or decrease welfare due to multiple frictions in the model. First, there is a moral hazard problem inside the firm due to information frictions. Second, there is monopolistic distortion in one of two sectors of this economy. Finally, there is a labor market friction due to random search. As a result, unemployment exists in labor submarkets. Therefore, a reduction in one friction need not increase welfare. It turns out that the factor governing the direction of the change in welfare is the elasticity

[31] The span of control is a constant which is not affected by the quality of MT and the demand draw in the continuous case. Readers are referred to the online appendix for details.

[32] In Bloom, Sadun, and Van Reenen (2012b), they find that a higher management score has a bigger positive impact on firm size in the U.S. than in other countries (e.g., column three of Table 4). As management scores are highly correlated with firm productivity, this implies that more productive firms gain more in economies with better MT like the U.S., which is exactly the key prediction of my model.

[33] Detailed discussions are available upon request.

Figure 6: Gains in Aggregate Productivity



of substitution between products in the CES sector, as it determines whether or not the CES sector expands after an improvement in MT.

Workers in the CES sector face a trade-off between lower wage and the higher probability of being employed. When MT improves, employed workers receive lower wages and payoffs (i.e., wages minus the disutility to exert effort) on average. However, firms expand and demand more labor due to better MT.[34] On top of that, the elasticity of substitution determines the sensitivity of the firm's expansion (i.e., increase in employment) with respect to an improvement in MT. When products are more substitutable, this sensitivity is higher. Thus, the increase in employment *per firm* is bigger. Moreover, this bigger increase in employment per firm eventually increases the aggregate income of the economy which makes the market size bigger. As a result, the CES sector accommodates more firms and its aggregate labor demand increases, which reduces the risk of being unemployed for workers in the CES sector. In summary, the increase in the probability of being employed dominates wage loss, which pushes up the workers' expected payoff from entering the CES sector when MT improves and the elasticity of substitution is high.

The opposite scenario takes place for the workers' expected payoff from entering the CES sector when the elasticity of substitution is lower. In this case, the increase in employment per firm is small when MT improves. This small increase in employment per firm and the decrease in the average wage eventually push down the aggregate income of the economy, which makes the market size smaller. As a result, the CES sector accommodates fewer firms and its aggregate labor demand *decreases*. Therefore, the

[34]Remember that the average firm size in terms of employment increases when MT improves.

worker obtains lower expected payoff from entering the CES sector, as both the average wage and the average employment rate in the CES sector decrease when the elasticity of substitution is lower. Note that although firms with the worst demand draws are driven out of the market when MT improves, the ideal price index increases. This is because the decrease in the mass of firms dominates the decrease in the average price charged by active firms. In total, welfare decreases when the elasticity of substitution is sufficiently small.

Note that the welfare implications here contrast with the implications in other models using CES preferences (e.g., Melitz (2003) and Caliendo and Rossi-Hansberg (2012)). In their models, an economy-wide improvement in firms' efficiency results in higher real wages (or welfare) unambiguously. With improvement in efficiency due to trade liberalization, firms demand more labor. As the wage is fully flexible, labor is fully employed both before and after the change. Therefore, the only way to counterbalance the increased labor demand is to raise the real wage, which results in higher welfare. However, as there are unemployed workers in the labor submarkets of the CES sector in equilibrium, better MT actually pushes down the average wage of employed workers in the CES sector in my model. Thus, whether or not the workers' (ex ante) expected payoff increases depends on whether or not there is a substantial decrease in the average unemployment rate. It turns out that the substantial decrease in the average unemployment rate is present only when the elasticity of substitution is big enough, as both the employment per firm and the mass of firms increase in this case.

Although the above discussion on welfare is a conjecture, simulation results do show that welfare can either increase or decrease after an improvement in MT. Table 1 presents an example in which welfare increases when MT improves, while Table 2 shows an example in which welfare decreases when MT improves. The online appendix shows that aggregate-level predictions of improvements in MT (e.g., changes in the exit cutoff, the firm size distribution, the distribution of the number of layers, and aggregate productivity) in the case of the continuous number of layers are qualitatively the same as those predictions in the case of the discrete number of layers. Thus, I derive and discuss analytical results on welfare by treating the number of layers as a continuous variable in the online appendix. I show that welfare increases after an improvement in MT if and only if $\frac{\sigma-1}{\sigma} > \gamma$. Due to the same reason, I treat the number of layers as a continuous variable to derive analytical results on the welfare gains from trade in the online appendix as well.

Table 1: Change in Welfare when MT Improves and σ is Big

	Welfare	Ave(wage)	Ave(ur)	M	E
b=1.6	0.2873	0.9582	0.5644	1.0235	25.6156
b=1.5	0.3492	0.8988	0.4178	1.1471	31.3347

ur: unemployment rate; M: the mass of active firms; E: total income

$\sigma = 3.8$, $\gamma = 0.6$, $\psi = 0.3$

3.2 Two-Dimension Heterogeneity in Management Quality

The data from the World Management Survey (WMS) show that there is substantial heterogeneity in the quality of MT both across countries and across firms. Thus, I

Table 2: Change in Welfare when MT Improves and σ is Small

	Welfare	Ave(wage)	Ave(ur)	M	E
b=1.6	0.3819	0.9530	0.4145	2.6066	53.4884
b=1.5	0.2910	0.8956	0.5119	1.9007	41.0722

ur: unemployment rate; M: the mass of active firms; E: total income
$\sigma = 2.8$, $\gamma = 0.75$, $\psi = 0.3$

now consider the case in which MT differs both across economies and across firms. More specifically, I assume that firms receive both the demand draw (i.e., θ) and the inefficiency of MT (i.e., b) after entry, and these two draws are independent. Under these assumptions, I show that results established in Proposition 8 hold in this case as well.[35] The detailed proof is relegated to the online appendix, as it is similar to the proof of Proposition 8. Moreover, because of variations in MT across firms, the model yields the following testable prediction which will be tested in Section 5.

Proposition 9 *Suppose management technology differs both across economies and across firms. An improvement in a firm's own management technology increases its firm size, the number of layers (weakly), and the span of control given the number of layers. However, an improvement in all other firms' management technology holding constant the firm's own management technology reduces the firm's size.*

Proof. See Appendix 7.2.10.

The key to understanding the above testable prediction is the pro-competitive effect of an improvement in MT. On the one hand, when MT improves only for a single firm, its firm size and profit unambiguously increase as better MT reduces labor costs. Furthermore, the firm becomes more decentralized by increasing either the span of control or the number of layers. On the other hand, when MT improves on average for all firms except for one firm, that firm's revenue and profit shrink because of the intensified competition. In total, variation in the quality of MT across firms and economies yields testable predictions that can be contrasted with the data.

4 Trade Liberalization and the Welfare Gains from Trade

Most countries are trading with each other, and trade liberalization brings changes to the internal organization of firms. Guadalupe and Wulf (2010) show that American firms in sectors with larger reductions in U.S. import tariffs following the enactment of the Canada-U.S. free trade agreement (FTA) flattened their hierarchies. They did so by reducing the number of layers between the chief executive officer (CEO) and division managers and increasing the span of control of the CEO. Furthermore, division managers

[35]Now, there are one FE condition and n ZCP conditions, if there are n possible outcomes of the draw of b. More specifically, the exit cutoff for the quality draw depends on both the adjusted market size A and the draw of b.

received more incentive-based pay after the CEO had increased the span of control. In order to rationalize these findings, I extend the baseline model to an international context. In this section, I maintain the assumption that the quality of MT differs only across countries.

4.1 Trade Liberalization and the Internal Organization of Firms

This subsection focuses on how firms respond to trade liberalization. More specifically, I am interested in how the internal organization of firms evolve after the economy opens up to trade. My analysis of opening up to trade in the symmetric two-country case follows Melitz (2003). I make the standard assumption that there are a fixed trade cost denoted by f_x and a variable trade cost denoted by $\tau(\geq 1)$ for firms in the CES sector to export. Similar to the fixed production cost, the fixed trade cost is also paid in the form of the final composite good defined in equation (1). The variable trade cost means that if τ units of output are shipped to the foreign country, only one unit arrives. Furthermore, it is assumed that the fixed trade cost is big enough that there is selection into exporting in the CES sector. The homogeneous good is not traded regardless of trade costs, because the two countries are symmetric.

In a world consisting of two countries, the firm in the CES sector allocates output between the two markets to equalize its marginal revenues. The optimal allocation of output in the domestic market is

$$q_d = \frac{q\left(\frac{A_H}{A_F}\tau^{\beta}\right)^{\sigma}}{1+\left(\frac{A_H}{A_F}\tau^{\beta}\right)^{\sigma}}, \tag{32}$$

where q is the total output, and A_H and $A_F(=A_H)$ are the adjusted market sizes of the domestic market and the foreign market, respectively. For non-exporters, the adjusted market size is A_H. For exporters, the adjusted market size is

$$\left(1+\frac{1}{\tau^{\sigma-1}}\right)^{\frac{1}{\sigma}}A, \tag{33}$$

where $A \equiv A_H = A_F$.

The equilibrium conditions in the open economy are similar to those in the closed economy except for the following modifications. First, an equilibrium condition that pins down the cutoff for exporting (i.e., $\bar{\theta}_x$) appears as

$$\Pi(\bar{\theta}_x, (1+\frac{1}{\tau^{\sigma-1}})^{\frac{1}{\sigma}}A) - \Pi(\bar{\theta}_x, A) = f_x. \tag{34}$$

Note that the cutoff for exporting cannot be solved analytically as the average cost is endogenously determined and depends on the number of layers the firm has. Second, the FE condition in equation (23) now becomes

$$\int_{\bar{\theta}}^{\bar{\theta}_x} \Pi(\theta, A)g(\theta)d\theta + \int_{\bar{\theta}_x}^{\infty} \Pi(\theta, (1+\frac{1}{\tau^{\sigma-1}})^{\frac{1}{\sigma}}A)g(\theta)d\theta = f_e, \tag{35}$$

as non-exporters and exporters face different adjusted market sizes. Third, the labor demand of firms in the CES sector (i.e., equation (26)) is also affected by the trade

costs in the open economy, and it includes the labor demand of both non-exporters and exporters. Finally, the market-clearing condition of the final composite good in equation (29) has to be modified as

$$\begin{aligned} E &= \int_{\bar{\theta}}^{\bar{\theta}_x} LC(\theta, A)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta + \int_{\bar{\theta}_x}^{\infty} LC(\theta, (1+\frac{1}{\tau^{\sigma-1}})^{\frac{1}{\sigma}}A)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta \\ &\quad \Big[f_0 + f_x\big(\frac{\bar{\theta}}{\bar{\theta}_x}\big)^k + f_e\big(\frac{\bar{\theta}}{\theta_{min}}\big)^k\Big]M + \psi M, \end{aligned} \tag{36}$$

as exporting firms use the final composite good to pay for the fixed trade cost and face a different market size compared with exporters. The equilibrium in the open economy is characterized by seven equations (i.e., equations (22), (24), (26), (28), (34), (35), and (36)) and seven endogenous variables (i.e., $\bar{\theta}$, $\bar{\theta}_x$, M, p_h, L_c, L_h and E). Using the same approach as the one used in the closed economy, I can prove that there exists a unique equilibrium characterized by (A, $\bar{\theta}$, $\bar{\theta}_x$ and p_h) under some restrictions on parameters' values. I derive values for all other endogenous variables based on these four variables.

I analyze how opening up to trade affects the internal organization of firms. The key to understanding why opening up to trade brings about differential impact on non-exporters and exporters is the difference in the change of the adjusted market size. The following lemma shows that the adjusted market size shrinks for non-exporters and increases for exporters.

Lemma 2 *When the economy opens up to trade, the adjusted market size faced by non-exporters decreases, while the adjusted market size faced by exporters increases. Furthermore, the exit cutoff for the quality draw increases.*

Proof: See Appendix 7.2.11.

With Lemma 2 in hand, I can analyze how the internal organization of firms and firm productivity change when the economy moves from autarky to the open economy. The main result is that non-exports flatten their hierarchies by reducing the number of layers and increasing the span of control, while exporting firms increase the number of layers and reduce the span of control. The following proposition summarizes these results.

Proposition 10 *When the economy opens up to trade, non-exporting firms reduce firm size, while exporting firms increase firm size. Non-exporting firms de-layer weakly, and increase the span of control when the number of layers is reduced. Exporting firms increase the number of layers weakly, and reduce the span of control when a new layer is added. Non-exporters increase the amount of incentive-based pay when they de-layer.*

Proof: See Appendix 7.2.12.

The effect of a bilateral trade liberalization on firms' internal organization is heterogeneous and depends on the situation the firm faces. In the theory, non-exporting firms reduce the number of layers and increase the span of control due to the shrinking market size after a bilateral trade liberalization. This is what Guadalupe and Wulf (2010) find for American firms located in industries with increasing import competition. Furthermore, according to the same theory, firms increase the incentive-based pay for employees

owning to the increasing span of control. This is another finding from Guadalupe and Wulf (2010). Of course, the theory also predicts that firms with increasing opportunities to export (weakly) increase their number of layers and reduce the span of control after a bilateral trade liberalization. Guadalupe and Wulf (2010) do find that American firms in sectors with larger reductions in Canadian import tariffs increased the number of layers and reduced the span of control, although these results are not statistically significant. In sum, firms facing different changes in the market environment adjust their internal organization differently after a bilateral trade liberalization.

4.2 Management Technology and the Welfare Gains from Trade

The quality of MT matters for how much an economy benefits from opening up to trade. I find that an economy with superior MT benefits disproportionately more from opening up to trade.[36] The intuition is firms with better demand draws receive higher profits when MT improves. These firms export and expand when the economy opens up to trade as well. Therefore, an economy with better MT makes the improvement in the organization of exporters more valuable. This result suggests that countries with poor institutions that hamper monitoring are likely to gain less from participation in the world economy.

I treat the number of layers as a continuous variable and derive analytical results on the welfare gains from trade (WGT) in the online appendix. As the model has multiple frictions, welfare losses from trade can in principle arise. I characterize a necessary and sufficient condition for the existence of the WGT in the online appendix. Furthermore, I derive a sufficient condition for the existence of the WGT and prove the existence of a complementarity between the WGT and the quality of MT under this sufficient condition. The sufficient condition implies that an economy gains from opening up to trade if the elasticity of substitution is large and the quality of MT is high. As the elasticity of substitution is different between the homogeneous sector and the CES sector, the WGT arises when the difference in it between two sectors is not too big (i.e., σ is big enough). Readers are referred to the online appendix for more details.

5 Evidence

In this section, I present econometric evidence to confirm the model's predictions for firm size using the data sets from the WMS.[37]

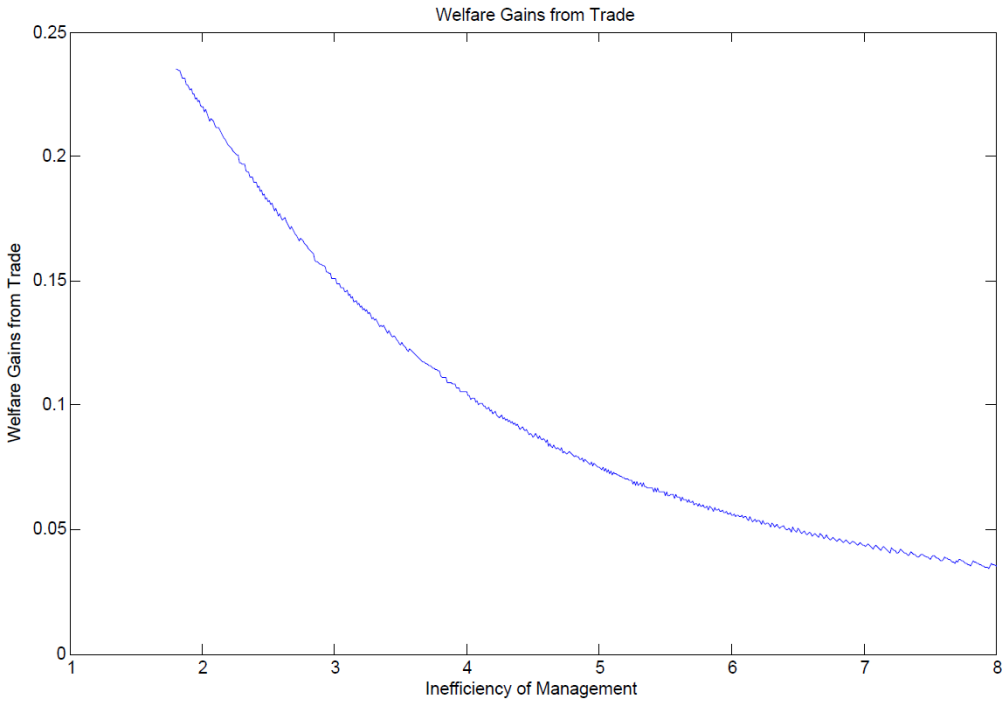
5.1 Data

Two main data sets I use were obtained from the WMS, and they were originally used in Bloom and Van Reenen (2010). The first data set is a cross-sectional data set that contains roughly 5700 firms across 16 countries. Each firm was interviewed once between

[36]Note that the complementarity between the welfare gains from trade and the quality of MT only holds *globally*, as Figure 7 shows. As firms *discontinuously* change their output, employment and prices when adjusting the number of layers, the complementarily does not hold locally. When the number of layers is treated as a continuous variable, the complementarity property holds both globally and locally. For more details, see the online appendix.

[37]For details of the empirical work, see both Appendix 7.1 and Appendix 7.3.

Figure 7: The Welfare Gains from Trade and Management Technology



2004 and 2008, and the score on each management practice was given after the firm had been interviewed. The second data set is a panel data set that contains accounting information (e.g., employment, assets and sales etc.) for firms whose management scores were given in the first data set. The time span for the second data set is from 2003 to 2008. I obtained country-level data from the website of either the World Bank or the Penn World Table. In particular, the labor market rigidity index that measures the easiness of firing workers is included in the first data set I obtained from the WMS and is also available from the website of the World Bank.

I constructed a variable called *moinc* to measure the quality of MT to monitor and incentivize workers. There are eighteen management practices defined in Bloom and Van Reenen (2010) in total, and these practices are grouped into three categories: monitoring, targets and incentives. I picked up seven management practices from them and argue that these seven management practices are closely related to the quality of MT to monitor and incentivize workers. For instance, management practices such as "performance tracking" and "performance review" are related to whether or not the firm can successfully find or catch misbehaving employees. Management practices such as "rewarding high performance" and "remove poor performers" are related to whether or not the firm can credibly reward hard-working employees and punish shirking workers. I calculated the average score on the seven management practices (i.e., *moinc*) and treat it as a measure for the quality of MT.

I used the Olley-Pakes approach to estimate firms' total factor productivity (TFP). One complication is that I cannot estimate the production function using all firms in a given industry, as firms that have different numbers of layers have different production functions. One solution is to obtain information on the number of layers for each firm and

group firms into several bins based on the number of layers they have. Unfortunately, such information is not available in the data sets I obtained from the WMS. I circumvented this problem by using firm size (i.e., deflated sales or employment) as a *proxy* for the number of layers, as the theory predicts that bigger firms have more layers. More specifically, I grouped firms in the same sector into two bins based on their deflated sales or employment and estimated the production function for firms in each bin separately using the Olley-Pakes approach.

5.2 Specification

I want to confirm two predictions of Proposition 9. First, the management score on monitoring and incentives at the firm level (i.e., *moinc*) positively affects firm size, ceteris paribus. Second, the average quality of MT of firms in an economy *negatively* affects a firm's size, ceteris paribus. This is the general-equilibrium effect of an improvement in MT. I use two variables to measure the average quality of MT in an economy. First, the average value of *moinc* across firms within an industry is used to capture the average quality of MT in an industry. Second, as the labor market rigidity index negatively affects firms' ability to incentivize workers, I use it as a proxy for the average quality of MT in a country.[38]

More specifically, I run two regressions. I consider the following specification first.

$$\begin{aligned} \log(size_{it}) &= \beta_0 + \beta_1 \log(TFP)_{it} + \beta_2\, moinc_{it} + \beta_3\, rigidity_c \\ &\quad + yearind_{tj} + controls_{ct} + \epsilon_{it}, \end{aligned} \tag{37}$$

where i is the firm ID; j indicates the industry the firm belongs to; c represents the country the firm is from; and t is the year. For the left-hand-side variable, I use both employment and deflated sales as measures for firm size. For the right-hand-side variables, $\log(TFP)_{it}$ captures variation in any firm-level variable that affects firm productivity; $moinc_{it}$ captures variation in the quality of MT; $yearind_{tj}$ (i.e., the year-industry fixed effect) captures any shock that is specific to an industry, but common across countries in a given year (e.g., drop in the world price of steel for the steel industry in a given year). I cannot include the country fixed effect into the regression since the labor market rigidity index does not vary across years. Instead, I include a set of country-level controls that affect firm size not through affecting the quality of MT into the regression. $\log(GDP)$ is used to capture the market size effect; $\log(GDPpercapita)$ is used to capture variation in the quality of physical capital, human capital, and scientific technology across countries; and the Doing Business index is used to capture variation in the quality of political and financial institutions across countries. For this regression, the theory predicts that both the TFP and the management score (i.e., $moinc_{it}$) at the firm level positively affects firm size, ceteris paribus. On top of that, the labor market rigidity index should *positively* affect firm size after I have controlled for firm-level characteristics as well as country-level controls.

[38] Based on the theory, what I need here are *exogenous* variations in the average quality of MT from the perspective of the firm. As the labor market rigidity index is probably out of a single firm's control, the variation in it is what I need. As it is hard to find an exogenous variable that affects the average quality of MT across firm in an industry, I treat the variation in the average value of *moinc* across firms within an industry as the variation in the average quality of MT.

Second, I run a regression of firm size on firm-level characteristics and the average score of *moinc* . More specifically, the specification is

$$\begin{aligned}\log(size_{it}) &= \beta_0 + \beta_1 \log(TFP)_{it} + \beta_2\, moinc_{it} + \beta_3\, ave\log(TFP)_{jct} \\ &\quad +\beta_4\, avemoinc_{jct} + yearctry_{tc} + yearind_{tj} + \epsilon_{it}.\end{aligned} \tag{38}$$

For the right-hand-side variables, $yearctry_{tc}$ (i.e., the year-country fixed effect) captures any macro-level shock that is specific to a country, but common across industries in a given year (e.g., an economic recession happening in the U.S. in a given year); $ave\log(TFP)_{jct}$ (i.e., the weighted average of $\log(TFP)$) captures variations in the overall market competitiveness; and $avemoinc_{jct}$ (i.e., the weighted average of *moinc*) captures variations in the market competitiveness due to the change in the average quality of MT. Similarly, the theory predicts that both the TFP and the management score positively affect firm size, ceteris paribus. Moreover, the average management score should negatively affect firm size, ceteris paribus.[39]

5.3 Results

Regression results reported in Tables 3 and 4 confirm the theoretical results established in Proposition 9. First, the coefficient in front of *moinc* is positively significant in all regressions. This is consistent with the prediction of the effect of the individual firm's management score on firm size. Second, the coefficient in front of the labor market rigidity index is positively significant in most regressions reported in Table 3, even when I control for a variety of country-level variables. This is consistent with the theoretical prediction that firm size is smaller in countries with less rigid labor market conditional on firm-level characteristics. Third, the coefficient in front of the average management score is negatively significant in all regressions reported in Table 4, even when I control for the weighted average of $\log(TFP)$. This result is again consistent with my theoretical prediction. Namely, firm size is smaller in an economy that has better managed firms conditional on firm-level efficiencies. Finally, the impact of the average quality of MT on firm size is quantitatively important. For instance, an increase in the average management score from 2.65 (i.e., India) to 3.33 (i.e., the U.S.) leads to about 7.5% decrease in a firm's size, ceteris paribus. In summary, the model gains support from the data.

Finally, I did several robustness checks for results reported in Tables 3 and 4. First, I re-estimated the TFP by grouping firms in each sector into *three* bins and implementing the Olley-Pakes productivity estimation. After having obtained the new TFP estimates, I rerun the regressions specified in equations (37) and (38) and found that regression results are qualitatively similar to what are reported in Tables 3 and 4. Second, I run the regressions specified in equations (37) and (38) for firms in each bin separately, as the effect of the quality of MT on firm size might be heterogeneous across firms that have different numbers of layers. The estimation results for both groups of firms are consistent with the predictions in Proposition 9 as well. For more details on the robustness checks, see Appendix 7.3.

[39] One potential concern here is that firms might not be *randomly* selected for the survey. The description of the data set in Bloom and Van Reenen (2010) shows that they randomly sampled medium-sized firms, employing between 100 and 5,000 workers.

Table 3: Labor Market Rigidity, Management Scores, and Firm Size

	(1) log(*employment*)	(2) log(*sales*)	(3) log(*employment*)	(4) log(*sales*)
log($TFP1$)	0.140***	1.007***		
	(0.0404)	(0.0342)		
log($TFP2$)			0.0180	0.146***
			(0.0220)	(0.0286)
moinc	0.223***	0.253***	0.236***	0.347***
	(0.0271)	(0.0286)	(0.0271)	(0.0354)
$Rigidityofemployment$	0.00625**	0.00424	0.00669***	0.00790**
	(0.00255)	(0.00273)	(0.00252)	(0.00319)
Constant	8.231***	6.848***	7.799***	3.745***
	(1.082)	(0.965)	(1.078)	(1.264)
MNE dummies	Yes	Yes	Yes	Yes
Industry-year F.E.	Yes	Yes	Yes	Yes
N	3329	3329	3339	3339
R^2	0.269	0.671	0.258	0.436
adj. R^2	0.208	0.644	0.197	0.389

$TFP1$: estimated TFP when firms are grouped into two bins based on employment.

$TFP2$: estimated TFP when firms are grouped into two bins based on deflated sales.

Country-level controls: log(GDP), log($GDPpercapita$), log($priceindex$) and the Doing Business index.

Standard errors are in parentheses and clustered at the country-industry level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Average Management Score and Firm Size: Industry-Level Analysis

	(1) log(*employment*)	(2) log(*sales*)	(3) log(*employment*)	(4) log(*sales*)
log($TFP1$)	0.129***	1.001***		
	(0.0411)	(0.0349)		
log($TFP2$)			0.0151	0.156***
			(0.0240)	(0.0332)
moinc	0.247***	0.272***	0.258***	0.363***
	(0.0318)	(0.0333)	(0.0321)	(0.0409)
ave log($TFP1$)	-0.00679	-0.0208		
	(0.0292)	(0.0260)		
ave log($TFP2$)			0.0224	-0.0123
			(0.0252)	(0.0283)
ave(*moinc*)	-0.116*	-0.116**	-0.117*	-0.131*
	(0.0607)	(0.0580)	(0.0608)	(0.0752)
Constant	5.950***	5.094***	6.302***	9.381***
	(0.371)	(0.349)	(0.352)	(0.451)
MNE dummies	Yes	Yes	Yes	Yes
Industry-year F.E.	Yes	Yes	Yes	Yes
Country-year F.E.	Yes	Yes	Yes	Yes
N	3113	3113	3121	3121
R^2	0.288	0.686	0.280	0.456
adj. R^2	0.228	0.660	0.219	0.410

$TFP1$ estimated TFP when firms are grouped into two bins based on employment.
$TFP2$ estimated TFP when firms are grouped into two bins based on deflated sales.
ave(A): weighted average of variable A at the industry-country-year level.
Firms in industry-country-year pairs that contain only one observation are excluded.
Standard errors are in parentheses and clustered at the country-industry level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6 Conclusions

This paper uses one canonical approach to modeling the incentive problem inside the firm and incorporates an incentive-based hierarchy into a general equilibrium framework to show the pro-competitive effect of an improvement in MT. By investigating how the quality of MT affects firm characteristics, this paper rationalizes several key findings in the macro-development literature and the organizational economics literature. On top of that, by extending the baseline model into the international context, I not only can explain several interesting findings related to a bilateral trade liberalization, but can also discuss how the quality of MT affects the welfare gains from trade.

The key role of improved MT is to intensify market competition and make more efficient firms obtain bigger market shares. As a result, firms are smaller and less decentralized on average in countries whose firms have worse MT. These key findings established in the empirical literature can be explained by my model. Furthermore, this paper can rationalize why firms flatten hierarchies and increase the use of incentive-based pay for employees when facing increasing import competition. Finally, the quality of MT affects the response of the economy to trade liberalization. Significantly, better MT increases the welfare gains from trade.

Undoubtedly, much more research remains to be done. First, it is extremely important to use the current framework to quantitatively evaluate how the quality of MT affects average productivity of firms in the economy. Second, integrating the knowledge-based hierarchy and the incentive-based hierarchy into a unified framework is an interesting idea, as each approach reflects only one part of the function of the hierarchy. Finally, investigating how differences in MT across countries affect trade patterns is worth doing as well.

7 Appendix

7.1 Empirical Motivation

In this subsection, I discuss details of my data work. Three parts come in order. First, I discuss the content of MT used in this paper. Next, I argue that MT affects firm performance by showing some motivating evidence.[40] Third, I show that quality of monitoring and incentives is an important part of the overall management quality, and there is substantial heterogeneity on these management practices across firms. Finally, I argue that the quality of MT differs across economies and is systematically correlated the firm size distribution.

First, I discuss what MT used in this paper means. In Bloom and Van Reenen (2010), eighteen management practices are grouped into three categories; monitoring, targets and incentives. "Monitoring" refers to "how companies monitor what goes on inside their firms and use this for continuous improvement". "Targets" refers to "how companies set the right targets, track the right outcomes, and take appropriate action if the two are inconsistent". "Incentives" refers to "how companies promote and reward employees based on performance, and whether or not companies try to hire and keep their best employees". This paper focuses on the first type and a part of the third type of management practices defined in Bloom and Van Reenen (2010).

I pick up seven management practices and argue that they are closely related to MT discussed in this paper. Among the seven practices I pick up, the first four items which are "performance tracking", "performance review", "performance dialogue" and "performance clarity" are related to whether or not the firm can successfully find and catch misbehaving employees. The other three items which are "consequence management", "rewarding high performance", "remove poor performers" are related to whether or not the firm can credibly punish (and reward) shirking employees (and hard working employees). I calculate the average score on these seven items and treat it as the measure for the quality of MT. The average score on these seven items is defined as *moinc*.

Next, I show that the quality of MT affects firm performance substantially by presenting simple scatter plots in Figure 8. As it is evident in the figure, the quality of MT is positively associated with firm performance such sales per employee or total employment.

Third, I implement the variance and covariance decomposition exercise and show that there is substantial heterogeneity for scores on monitoring and incentives across firms. First, note that the average management score is the sum of two parts or

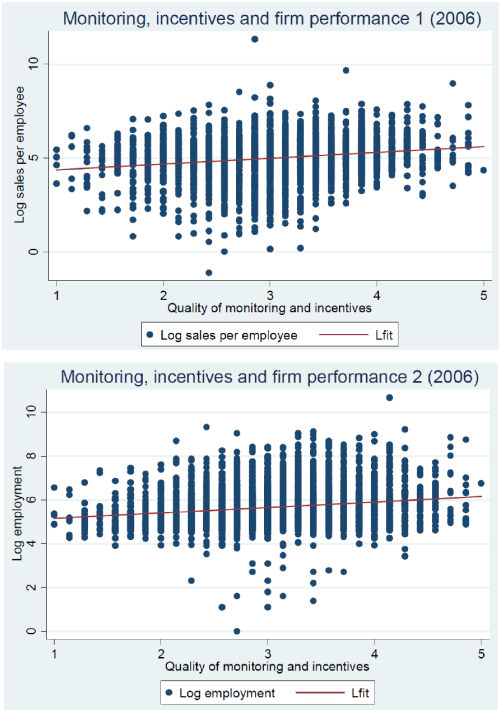
$$ms_i = moinc_i + nonmoinc_i = \frac{1}{18}\Sigma_{j\in moinc} Score_{ij} + \frac{1}{18}\Sigma_{j\in nonmoinc} Score_{ij},$$

where ms_i is the average management score for firm i, and j is the $j-$th management practice. The set *moinc* (or *nonmoinc*) is the set of management practices that are (or are not) related to monitoring and incentives. Second, I decompose the variation in the average management score into the following three parts:

$$\begin{aligned}\frac{1}{n}\Sigma_i(ms_i - \overline{ms})^2 &= \frac{1}{n}\Big[\Sigma_i(moinc_i - \overline{moinc})^2 + \Sigma_i(nonmoinc_i - \overline{nonmoinc})^2 \\ &\quad +2\Sigma_i(moinc_i - \overline{moinc})(nonmoinc_i - \overline{nonmoinc})\Big],\end{aligned}$$

[40] Rigorous empirical analysis is in Subsection 5.

Figure 8: Management Quality and Firm Performance



where $\bar{ms} = \frac{1}{n}\Sigma_i ms_i$ and n is the number of firms in the data set. The first and the second terms above are the variations in management scores coming from *moinc* and *nonmoinc* respectively. The last term reflects the correlation between these two scores across firms. Table 5 shows that there is substantial variation in the score of *moinc* across firms.

Table 5: Management Score Decomposition

	overall variation	var(moinc)	var(nonmoinc)	cross term
Contributions	0.442	0.071	0.187	0.184

moinc: the average score on 7 monitoring and incentive practices.

nonmoinc: the average score on 11 remaining measures.

crossterm: the correlation between the above two scores across firms.

Finally, I show that quality of monitoring and incentives differs across countries and is associated with the firm size distribution. First, Table 6 shows that the score on every management practice that is included in the set of *moinc* differs significantly between China and the U.S. Second, Figure 9 shows that the distribution of the score on *moinc* differs substantially between India and the U.S.[41] Namely, the average score on *moinc* is significantly higher for U.S. firms. Finally, the average management score on *moinc* is positively associated with the average firm size in an economy as Figure 9 shows. In other words, U.S. firms that have good MT seem to be larger than Indian firms on average, and there are much more small firms in India compared with the U.S.

[41] The upper graph uses data from the WMS. The lower graph comes from Hsieh and Klenow (2012).

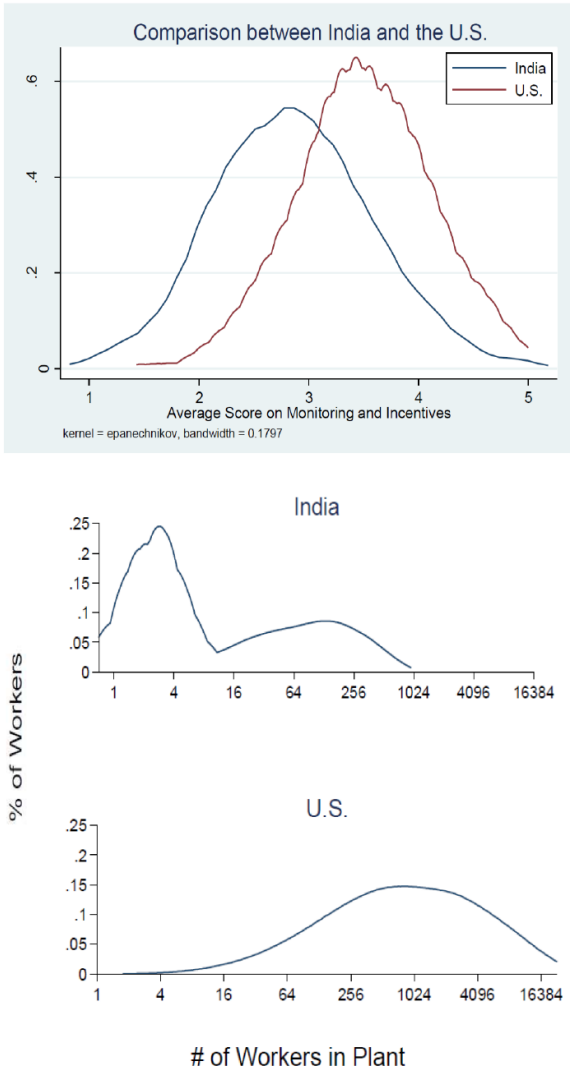
Table 6: Difference in the Quality of Monitoring and Incentives between China and the U.S.

	Perf2	Perf3	Perf4	Perf10	Perf5	talent2	talent3
U.S.	3.65	3.63	3.58	2.93	3.67	3.17	3.82
China	2.99	3.09	2.74	2.77	2.65	2.88	3.04

Range for scores: $1-5$.

Differences in means are all statistically significant at 1% level.

Figure 9: Management Quality and the Firm Size Distribution



In total, evidence presented in this subsection explains why cross-country difference in the quality of MT is a natural candidate to explain cross-country difference in aggregate variables such as the firm size distribution.

7.2 Proofs

7.2.1 Proof of Lemma 1

Proof. First, suppose there is a worker at layer i who shirks in equilibrium (i.e., $a_i = 0$). If he is a production worker, removing him from the hierarchy does not affect the firm's output and (weakly) reduces labor cost, which means it is optimal to exclude him from the hierarchy. If he is a non-production worker, his direct subordinates at layer $i+1$ would shirk as a result of the absence of monitoring from above. Furthermore, all his direct and *indirect* subordinates would shirk as well. Similar as before, removing them from the hierarchy does not affect the firm's output and (weakly) reduces labor cost, which means excluding them from the hierarchy is always optimal. Thus, all workers are incentivized to work in equilibrium. Second, the reason why the firm wants to allocate the monitoring intensity evenly across workers is that the firm could reduce wage payment by doing so if the monitoring intensity were not equalized. More specifically, suppose there are two units of labor inputs (i.e., time) that are monitored under different monitoring intensities p_1 and p_2. As all workers are incentivized to work, the wage payment to these two units of labor input equals

$$b\psi\Big(\frac{1}{p_1}+\frac{1}{p_2}\Big).$$

However, the firm can reduce this wage payment by equalizing the monitoring intensity across these two units of labor inputs as

$$2b\psi\frac{1}{(p_1+p_2)/2} < b\psi\Big(\frac{1}{p_1}+\frac{1}{p_2}\Big)$$

for any $p_1 \neq p_2$. This means that the firm can induce the two units of labor inputs to work under a lower cost. Therefore, the firm's optimal choice is to equalize the monitoring intensity across workers at a given layers. QED.

7.2.2 Proof of Proposition 1

First of all, let me write down the expression for the Average Variable Cost (AVC) function and the Marginal Cost (MC) function given the number of layers:

$$AVC(q,T) = (2-\frac{1}{2^{T-1}})b\psi 2^{1-\frac{T}{2^T-1}}q^{\frac{1}{2^T-1}} \tag{39}$$

and

$$MC(q,T) = b\psi 2^{2-\frac{T}{2^T-1}}q^{\frac{1}{2^T-1}} = \frac{2^T}{2^T-1}AVC(q,T). \tag{40}$$

From the expression of these two cost functions, it is straightforward to see that both of them increase with output q given the number of layers $T+1$. Thus, the first part of the proposition has been proved.

Next, I discuss the overall shape of the AVC curve. Before the discussion, let me make the following notation for future use.

Definition 1 *Let q_T be the solution to the following equation:*

$$AVC(q_T, T) = AVC(q_T, T+1).$$

In other words, the AVC of using $T+1$ layers is equal to the AVC of using $T+2$ layers at output level q_T.

Now, I prove the following lemma which assures the monotonicity of q_T.

Lemma 3 *q_T increases in T.*

Proof. I can rewrite $AVC(q_T, T) = AVC(q_T, T+1)$ as

$$\frac{(2-\frac{1}{2^{T-1}})/2^{\frac{T}{2^T-1}}}{(2-\frac{1}{2^{(T+1)-1}})/2^{\frac{(T+1)}{2^{(T+1)}-1}}} q_T^{\frac{1}{2^T-1}-\frac{1}{2^{(T+1)}-1}} = 1.$$

Thus, the switching point q_T can be rewritten as

$$q_T = \Big[\frac{2^{T+1}-1}{2^{T+1}-2}\Big]^{\frac{(2^T-1)(2^{T+1}-1)}{2^T}} 2^{(T-1)+\frac{1}{2^T}} \equiv \Psi_1(T)\Psi_2(T).$$

Taking logs and calculating the first order derivative with respect to T yields the following result:

$$\frac{d[\ln(\Psi_1(T)) + \ln(\Psi_2(T))]}{dT} = \ln 2(2^{T+1} - 2^{-T}) \ln\Big(\frac{2^{T+1}-1}{2^{T+1}-2}\Big) - \ln 2 + \ln 2\Big[1 - \frac{\ln 2}{2^T}\Big].$$

Thus, the sign of the above expression depends on

$$Sign\Big(\frac{d[\ln(\Psi_1(T)) + \ln(\Psi_2(T))]}{dT}\Big) = Sign\Big((2^{2T+1}-1)\ln(\frac{2^{T+1}-1}{2^{T+1}-2}) - \ln 2\Big)$$

or

$$Sign\Big(\frac{d[\ln(\Psi_1(T)) + \ln(\Psi_2(T))]}{dT}\Big) = Sign\Big((2^{2T+1}-1)\ln(\frac{2^{T+1}-1}{2^{T}-1}) - 2^{2T+1}\ln 2\Big)$$

or

$$Sign\Big(\frac{d[\ln(\Psi_1(T)) + \ln(\Psi_2(T))]}{dT}\Big) = Sign\Big((1-2^{-(2T+1)})\ln(\frac{2^{T+1}-1}{2^{T}-1}) - \ln 2\Big).$$

I want to show that $(1-2^{-(2T+1)})\ln(\frac{2^{T+1}-1}{2^T-1}) - \ln 2$ decreases in T for $T \geq 1$. First, I have

$$\begin{aligned}\frac{d\Big[(1-2^{-(2T+1)})\ln(\frac{2^{T+1}-1}{2^T-1}) - \ln 2\Big]}{dT} &= \ln 2\Big[\ln(\frac{2^{T+1}-1}{2^T-1})\frac{1}{2^{2T}} - \big(1-\frac{1}{2^{2T+1}}\big)\frac{1}{2^{T+1}-3+\frac{1}{2^T}}\Big] \\ &\equiv \ln 2(K_1(T) - K_2(T)).\end{aligned}$$

Second, I prove that

$$K_1(T) - K_2(T) < 0$$

for all $T \geq 1$. I proceed in two steps. In the first step, calculation shows that $K_1(1) - K_2(1) < 0$. In the second step, it is straightforward to see that for any $T > 1$

$$K_1(T) < \frac{1}{2^{2(T-1)}} K_1(1)$$

and

$$K_2(T) > \frac{1}{2^{T-1}} K_2(1).$$

Thus, I have

$$K_1(T) - K_2(T) < K_1(1) - K_2(1) < 0$$

for all $T > 1$. Finally, due to the monotonicity of $(1 - 2^{-(2T+1)}) \ln(\frac{2^{T+1}-1}{2^T-1}) - \ln 2$ with respect to T that has just been proven, I conclude that

$$(1 - 2^{-(2T+1)}) \ln(\frac{2^{T+1}-1}{2^T-1}) - \ln 2 > \lim_{T\to\infty} (1 - 2^{-(2T+1)}) \ln(\frac{2^{T+1}-1}{2^T-1}) - \ln 2 = 0.$$

and

$$Sign\Big(\frac{d[\ln(\Psi_1(T)) + \ln(\Psi_2(T))]}{dT}\Big) = Sign\Big((1 - 2^{-(2T+1)}) \ln(\frac{2^{T+1}-1}{2^T-1}) - \ln 2\Big) > 0.$$

Therefore, q_T must be an increasing function of T. QED.

Having established the monotonicity of q_T in Lemma 3, I can characterize the overall shape of the AVC curve.

Lemma 4 *If the output produced in equilibrium $q \in [q_{T-1}, q_T)$, the optimal number of hierarchical layers is $T + 1$. At the output level q_T, the AVC curve kinks and its slope decreases discontinuously as the firms adds a layer. Finally, the switching point q_T does not depend on the firm's quality draw (i.e., θ), the inefficiency of MT (i.e., b) and the adjusted market size (i.e., A).*

Proof. I proceed the proof in the following several steps. First, note that at q_{T-1}, the slope of $AVC(q, T)$ is smaller than the slope of $AVC(q, T-1)$ as $AVC(q_{T-1}, T) = AVC(q_{T-1}, T-1)$. This prove the second part of this lemma. Second, due to this property, $AVC(q, T-1)$ is below $AVC(q, T)$ for $q < q_{T-1}$ and above $AVC(q, T)$ for $q > q_{T-1}$. Thus, $T+1$ layers is never chosen for $q < q_{T-1}$. Similarly, $T+1$ layers is never chosen for $q > q_T$ as $AVC(q, T)$ is above $AVC(q, T+1)$ for $q > q_T$. Third, as $AVC(q, T)$ is below $AVC(q, T-1)$ for $q > q_{T-1}$ and q_T increases in T, $AVC(q, T)$ is below $AVC(q, t)$ for all $t < T$ when $q > q_{T-1}$. Similarly, as $AVC(q, T+1)$ is above $AVC(q, T)$ for $q < q_T$ and q_T increases in T, $AVC(q, t)$ is above $AVC(q, T)$ for all $t > T$ when $q < q_T$. In total, $AVC(q, T)$ is below $AVC(q, t)$ for all $t \neq T$ when $q \in (q_{T-1}, q_T)$ which leads to the result that for $q \in (q_{T-1}, q_T)$, the optimal choice of layers is $T + 1$. Of course, when $q = q_{T-1}$, choosing either T layers or $T + 1$ layers is optimal. Finally, the third half of the above lemma follows from the expression of q_T directly. QED.

I prove the following claim that characterizes the overall shape of the MC curve.

Claim 1 *Given the number of layers $T+1$, the MC increases with output. The final MC curve is*

$$MC(q) = MC(q, T)$$

where $q \in [q_{T-1}, q_T)$. This cost increases in interval $[q_{T-1}, q_T)$ for all T and decreases discontinuously *at the point q_T.*

Proof. It is straightforward to see the first part of this proposition due to Lemma 4. The only thing that needs proof is the last part. First, it is straightforward to see that $MC(q, T)$ increases in q for a given T. Second, at q_T, I have

$$AVC(q_T, T) = AVC(q_T, T+1).$$

As

$$MC(q, T) = \frac{2^T}{2^T - 1} AVC(q, T),$$

it must be true that

$$MC(q_T, T) > MC(q_T, T+1).$$

The fall in the marginal cost when the firm adds a layer comes from the reorganization inside the firm. QED.

In sum, I proved Proposition 1 due to Lemma 4 and Claim 1. QED.

7.2.3 Proof of Proposition 2

The first part of this proposition is true because of the shape of the AVC curve shown in Proposition 1. I prove the second of this proposition in five steps. First, I define two demand thresholds for a given number of layers $T+1$ for future use.

Definition 2 *For the number of layers $T+1$, θ_{T1} is defined as the solution to*

$$MR(\theta_{T1}, q_T) = A\beta\theta_{T1}^{\frac{1}{\sigma}} q_T^{-\frac{1}{\sigma}} = MC(q_T, T+1) = b\psi 2^{2-\frac{T+1}{2^{T+1}-1}} q_T^{\frac{1}{2^{T+1}-1}}.$$

In other words, firms with the quality draw θ_{T1} have their marginal revenue (MR) curve intersect the MC curve of using $T+2$ layers at output level q_T. $\theta_{T3} (> \theta_{T1})$ is defined as the solution to

$$MR(\theta_{T3}, q_T) = A\beta\theta_{T3}^{\frac{1}{\sigma}} q_T^{-\frac{1}{\sigma}} = MC(q_T, T) = b\psi 2^{2-\frac{T}{2^T-1}} q_T^{\frac{1}{2^T-1}}.$$

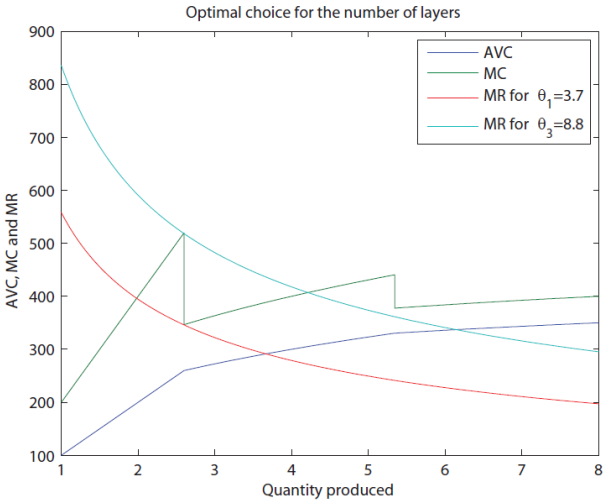
In other words, firms with the quality draw θ_{T3} have their MR curve intersect the MC curve of $T+1$ layers at output level q_T.

The graphical representation of θ_{T1} and θ_{T3} is in Figure 10.

Second, having defined the two demand thresholds for T, I show that only when the firm's quality draw is between $[\theta_{T1}, \theta_{T3}]$, does it have incentive to switch from having $T+1$ layers to having $T+2$ layers in Lemma 5.

Lemma 5 *For each T, firms having the quality draw smaller than or equal to θ_{T1} prefer $T+1$ layers over $T+2$ layers, while firms having the quality draw higher than or equal to θ_{T3} prefer $T+2$ layers over $T+1$ layers.*

Figure 10: Lower and Upper Bounds on Layer-Switching from $T = 1$ to $T = 2$



Proof. First, note that as $MC(q_T, T) > MC(q_T, T+1)$ and $MR(\theta, q)$ is an increasing function of θ for a given q, it must be true that $\theta_{T1} < \theta_{T3}$.

Next, if a firm with $\theta < \theta_{T1}$ chose $T+2$ layers, it must be true that $q(\theta, T+1) < q_T$ which is not optimal for the firm as $AVC(q, T) < AVC(q, T+1)$ for output levels smaller than q_T. Thus, Firms with $\theta < \theta_{T1}$ prefer $T+1$ layers over $T+2$ layers. Similarly, if a firm with $\theta > \theta_{T3}$ chose $T+1$ layers, it must be true that $q(\theta, T) > q_T$ which contradicts that $AVC(q, T) > AVC(q, T+1)$ for output levels bigger than q_T. Thus, Firms with $\theta > \theta_{T3}$ prefer $T+2$ layers over $T+1$ layers.

Finally, when $\theta = \theta_{T1}$, choosing $T+1$ layers yields more profit as

$$\pi(\theta_{T1}, T) \equiv \pi(\theta_{T1}, T, q(\theta_{T1}, T)) > \pi(\theta_{T1}, T, q_T) = \pi(\theta_{T1}, T+1, q_T),$$

where I have used the result that $AVC(q_T, T) = AVC(q_T, T+1)$. Similarly, when $\theta = \theta_{T3}$, choosing $T+2$ layers yields more profit as

$$\pi(\theta_{T3}, T+1) \equiv \pi(\theta_{T3}, T+1, q(\theta_{T3}, T+1)) > \pi(\theta_{T3}, T+1, q_T) = \pi(\theta_{T3}, T, q_T).$$

QED.

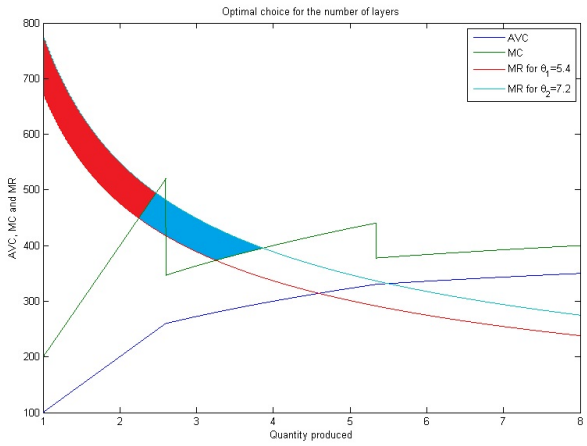
Third, I prove the following lemma showing the complementarity between the benefit of adding a layer and the quality draw θ.

Lemma 6 *For a given T, $\pi(\theta, T+1) - \pi(\theta, T)$ continuously increases in θ for $\theta \in [\theta_{T1}, \theta_{T3}]$.*

Proof. I use Figure 11 to prove this lemma. For any $\theta \in [\theta_{T1}, \theta_{T3}]$, suppose the quality draw θ increases by $\Delta > 0$ which corresponds to a shift of the MR curve from the red one to the green one. The difference between $\pi(\theta, T)$ and $\pi(\theta+\Delta, T)$ is represented by the red region, while the difference between $\pi(\theta, T+1)$ and $\pi(\theta+\Delta, T+1)$ is represented by the sum of the red region and the blue region. Thus, I have

$$\begin{aligned} & \pi(\theta+\Delta, T+1) - \pi(\theta+\Delta, T) - [\pi(\theta, T+1) - \pi(\theta, T)] \\ = \; & [\pi(\theta+\Delta, T+1) - \pi(\theta, T+1)] - [\pi(\theta+\Delta, T) - \pi(\theta, T)] \end{aligned}$$

Figure 11: Optimal Choice of the Number of Layers



which is the blue region. As the MR curve moves upward when θ increases and the MC curve of $T+2$ layers lies below the MC curve of $T+1$ layers when $q \geq q_{T1}$, the area of the blue region increases as Δ increases. Thus, it must be true that

$$\pi(\theta+\Delta, T+1) - \pi(\theta+\Delta, T) - [\pi(\theta, T+1) - \pi(\theta, T)]$$

increases in Δ which means that $\pi(\theta, T+1) - \pi(\theta, T)$ increases in θ for $\theta \in [\theta_{T1}, \theta_{T3}]$. The continuity of $\pi(\theta, T+1) - \pi(\theta, T)$ in θ is straightforward to see. QED.

Forth, I prove the following result which is the key step to prove this proposition. More specifically, there exists a threshold $\theta_{T2} \in (\theta_{T1}, \theta_{T3})$ such that firms with this level of efficiency is indifferent between having $T+1$ layers and having $T+2$ layers. Claim 2 summarizes the results.

Claim 2 *For each T, there exists a threshold $\theta_{T2} \in (\theta_{T1}, \theta_{T3})$ such that firms with this demand draw is indifferent between having $T+1$ layers and having $T+2$ layers. Moreover, firms with a level of the demand draw smaller than θ_{T2} strictly prefer $T+1$ layers over $T+2$ layers, while firms with a level of the demand draw bigger than θ_{T2} strictly prefer $T+2$ layers over $T+1$ layers.*

Proof. From Lemma 5, I have

$$\pi(\theta_{T1}, T) > \pi(\theta_{T1}, T+1),$$

and

$$\pi(\theta_{T3}, T) < \pi(\theta_{T3}, T+1).$$

As $\pi(\theta, T+1) - \pi(\theta, T)$ continuously increases in θ for $\theta \in [\theta_{T1}, \theta_{T3}]$ due to Lemma 6, there must exist a threshold $\theta_{T2} \in (\theta_{T1}, \theta_{T3})$ such that

$$\pi(\theta_{T2}, T) = \pi(\theta_{T2}, T+1).$$

And for all $\theta < \theta_{T2}$

$$\pi(\theta, T) > \pi(\theta, T+1),$$

while for all $\theta > \theta_{T2}$

$$\pi(\theta, T) < \pi(\theta, T+1).$$

QED.

Now, I can prove this proposition by generalizing Claim 2 into the case of any two different values of the number of layers. First, I define the upper bound and the lower bound on the quality draw for the firm's changing the number of layers from T_0 to $T_1(> T_0)$, where T_0 and T_1 can be any positive numbers. The following definition is used for this purpose.

Definition 3 *For the numbers of layers T_0 and $T_1(> T_0)$, θ_{T_0,T_1} is defined as the solution to*

$$MR(\theta 0_{T_0,T_1}, q_{T_0,T_1}) = MC(q_{T_0,T_1}, T_1),$$

where q_{T_0,T_1} is the output level at which $AVC(q_{T_0,T_1}, T_0) = AVC(q_{T_0,T_1}, T_1)$. $\theta 1_{T_0,T_1}(> \theta 0_{T_0,T_1})$ is defined as the solution to

$$MR(\theta 1_{T_0,T_1}, q_{T_0,T_1}) = MC(q_{T_0,T_1}, T_0).$$

Second, using the same approach used in the proof of Claim 2, one can prove that there exists a quality cutoff $\theta 2_{T_0,T_1} \in (\theta 0_{T_0,T_1}, \theta 1_{T_0,T_1})$ such that firms with quality draws bigger than $\theta 2_{T_0,T_1}$ prefer $T_1 + 1$ layers over $T_0 + 1$ layers can vice versa. Third, suppose there are two firms with quality draws θ_1 and $\theta_0 > (\theta_1)$ such that the firm with quality draw θ_0 has fewer layers than the firm with quality draw θ_1. I use $T_1 + 1$ and $T_0 + 1(< T_1 + 1)$ to denote the number of layers for firms with quality draws θ_1 and θ_0 respectively. Form the above discussion, it is straightforward to see that this supposition can't be true, as firms with quality draws bigger than $\theta 2_{T_0,T_1}$ prefer $T_1 + 1$ layers over $T_0 + 1$ layers and vice versa. Therefore, firms with better demand draws have more layers. QED.

Thanks to this proposition, I only need to derive the sequence of θ_{T2} for $T = 1, 2, 3...$ when solving the optimal number of layers for each firm. In other words, there is no need to solve the optimal number of layers for *each* firm respectively. Simulations I implement in this paper become much less time-consuming because of this result.

7.2.4 Proof of Proposition 3

First, when the firm expands without changing the number of layers, both employment and output increase continuously due to equations (12) and (14). Second, the firm's optimal pricing rule implies that

$$p(\theta) = \frac{\sigma}{\sigma - 1} MC(q(\theta)). \tag{41}$$

As the firm's MC increases given the number of layers and decreases discontinuously when the firm adds a layer due to a marginal increase in θ, the firm's price follows the same pattern. Finally, the firm's optimal output is

$$q(\theta) = \theta A^{\sigma} \Big(\frac{\sigma}{\sigma - 1} MC(q(\theta)) \Big)^{-\sigma}. \tag{42}$$

When the firm adds a layer due to a marginal increase in θ, the output increases discontinuously as the price falls discontinuously. The span of control increases at all existing layers as well which will be shown later. Therefore, employment jumps up *discontinuously* due to both the decreasing span of control at existing layers and the employed workers at the new layer. QED.

7.2.5 Proof of Proposition 4

First, equation (13) implies that the span of control increases at all layers when θ increases and its number of layers is unchanged. Second, as the wage defined in equation (8) is positively affected the span of control, wages increase at all layers. Third, the FOCs with respect to employment in equation (11) show that

$$\frac{w_i(q(\theta,T(\theta)),T(\theta))}{w_{i+1}(q(\theta,T(\theta)),T(\theta))}=\frac{1}{2}\frac{m_{i+1}(q(\theta,T(\theta)),T(\theta))}{m_i(q(\theta,T(\theta)),T(\theta))}=\frac{1}{2}x_i(q(\theta,T(\theta)),T(\theta))$$

for $T(\theta)>i\geq 1$. As the span of control increases at all layers, relative wages increase at all layers as well.

Finally, I prove the employment hierarchy that the number of workers is smaller in upper layers.[42] As I consider the employment hierarchy for workers, the minimum value for T is two. Equation (13) shows

$$\frac{m_{i+1}(q(\theta,T(\theta)),T(\theta))}{m_i(q(\theta,T(\theta)),T(\theta))}=2\Big[\frac{q(\theta,T(\theta))}{2^{T(\theta)}}\Big]^{\frac{2^{T(\theta)-(i+1)}}{2^{T(\theta)}-1}}\geq\Big[\frac{q(\theta,T(\theta))}{2^{T(\theta)-1}}\Big]^{\frac{2^{T(\theta)-(i+1)}}{2^{T(\theta)}-1}},$$

as $T(\theta)\geq 2$ and $T(\theta)>i\geq 1$. Now, I show the following property of q_{T-1} that is the key step to prove the result of the employment hierarchy:[43]

$$\frac{q_{T-1}}{2^{T-1}}=\Big[\frac{2^T-1}{2^T-2}\Big]^{\frac{(2^{T-1}-1)(2^T-1)}{2^{T-1}}}2^{\frac{1}{2^{T-1}}-1}>1.$$

This is because

$$\Big[\frac{2^T-1}{2^T-2}\Big]^{\frac{(2^{T-1}-1)(2^T-1)}{2^{T-1}}}2^{\frac{1}{2^{T-1}}-1}$$

increases in T for $T\geq 2$ and achieves its minimum value of 1.299 when $T=2$. In total,

$$\frac{m_{i+1}(q(\theta,T(\theta)),T(\theta))}{m_i(q(\theta,T(\theta)),T(\theta))}\geq\Big[\frac{q(\theta,T(\theta))}{2^{T(\theta)-1}}\Big]^{\frac{2^{T(\theta)-(i+1)}}{2^{T(\theta)}-1}}>\Big[\frac{q_{T(\theta)-1}}{2^{T(\theta)-1}}\Big]^{\frac{2^{T(\theta)-(i+1)}}{2^{T(\theta)}-1}}>1.$$

Therefore, the employment hierarchy holds for workers.

One implication of the employment hierarchy result is that the monitoring probability in firms with $T\geq 2$ is strictly smaller than one at every layer if $b\geq 1$, as the span of control is always smaller than one. QED.

[42] This result will be used in Appendix 7.2.8.

[43] q_T is defined in Definition 1.

7.2.6 Proof of Proposition 5

First of all, keep in mind that I am considering a small change in θ from $\theta_{T0,2} - \Delta$ to $\theta_{T0,2} + \Delta$ that triggers the addition of one layer into the hierarchy. Note that $\theta_{T0,2}$ is the demand threshold where the firm switches from having $T0 + 1$ layers to having $T0 + 2$ layers.

As the change in the span of control is the key to prove this proposition, I prove that the span of control falls at all existing layers first. From equations (12) and (14), I have

$$\frac{m_i^*}{m_{i-1}^*}\Big|_{T0} = 2\Big[\frac{\beta A(\theta_{T0,2} - \Delta)^{\frac{1}{\sigma}}}{4b\psi 2^{\frac{T0}{\sigma}}}\Big]^{\frac{\sigma 2^{T0-i}}{\sigma+(2^{T0}-1)}}.$$

and

$$\frac{m_{i+1}^*}{m_i^*}\Big|_{T0+1} = 2\Big[\frac{\beta A(\theta_{T0,2} + \Delta)^{\frac{1}{\sigma}}}{4b\psi 2^{\frac{T0+1}{\sigma}}}\Big]^{\frac{\sigma 2^{T0-i}}{\sigma+(2^{T0+1}-1)}},$$

where Δ is infinitesimally small. Thus, the thing I have to prove is that

$$Z(\theta_{T0,2}, T0) = \Big[\frac{\beta A\theta_{T0,2}^{\frac{1}{\sigma}}}{4b\psi 2^{\frac{T0}{\sigma}}}\Big]^{\frac{1}{\sigma+(2^{T0}-1)}}$$

decreases with $T0$ at $\theta_{T0,2}$. Calculation shows that

$$Sign\Big[dZ(\theta_{T0,2}, T0)/dT0\Big] = Sign\Big[\ln 2\Big[-2^{T0}\frac{(\ln\frac{\beta A\theta_{T0,2}^{\frac{1}{\sigma}}}{4b\psi} - \ln 2\frac{T0}{\sigma})}{2^{T0}+(\sigma-1)} - \frac{1}{\sigma}\Big]\Big].$$

Obviously, if

$$\frac{\beta A\theta_{T0,2}^{\frac{1}{\sigma}}}{4b\psi 2^{\frac{T0}{\sigma}}} \geq 1,$$

then the proof is done. So, I only need to consider the case where

$$\frac{\beta A\theta_{T0,2}^{\frac{1}{\sigma}}}{4b\psi 2^{\frac{T0}{\sigma}}} < 1.$$

For this case, there is a lower bound on the above term due to the result that $\theta_{T0,2} > \theta_{T0,1}$. Thus, I only have to prove that

$$-2^{T0}\frac{(\ln\frac{\beta A\theta_{T0,1}^{\frac{1}{\sigma}}}{4b\psi} - \ln 2\frac{T0}{\sigma})}{2^{T0}+(\sigma-1)} - \frac{1}{\sigma} < 0.$$

Based on Definition 2, $\theta_{T0,1}$ can be rewritten as

$$MR(\theta_{T0,1}, q_{T0}) = A\beta\theta_{T0,1}^{\frac{1}{\sigma}} q_{T0}^{-\frac{1}{\sigma}} = MC(q_{T0}, T0+1) = b\psi 2^{2-\frac{T0+1}{2^{T0+1}-1}} q_{T0}^{\frac{1}{2^{T0+1}-1}}.$$

Thus, I can solve $\theta_{T0,1}$ as

$$\theta_{T0,1} = \frac{(b\psi 2^{2-\frac{T0+1}{2^{T0+1}-1}})^{\sigma} q_{T0}^{\frac{\sigma+(2^{T0+1}-1)}{2^{T0+1}-1}}}{(A\beta)^{\sigma}}.$$

Consequently, I have

$$2^{T0}\frac{(\ln\frac{\beta A\theta_{T0,1}^{\frac{1}{\sigma}}}{4b\psi}-\ln 2\frac{T0}{\sigma})}{2^{T0}+(\sigma-1)}=\frac{2^{T0}(\sigma+(2^{T0+1}-1))}{(\sigma+(2^{T0}-1))\sigma(2^{T0+1}-1)}\ln\left(\frac{q_{T0}}{2^{T0}}\right)-\frac{2^{T0}}{(\sigma+(2^{T0}-1))(2^{T0+1}-1)}\ln 2.$$

As $\frac{q_{T0}}{2^{T0}}>1$ for $T0\geq 1$ due to the proof in Appendix 7.2.5, I conclude that

$$-2^{T0}\frac{(\ln\frac{\beta A\theta_{T0,1}^{\frac{1}{\sigma}}}{4b\psi}-\ln 2\frac{T0}{\sigma})}{2^{T0}+(\sigma-1)}-\frac{1}{\sigma}<\frac{2^{T0}}{(\sigma+(2^{T0}-1))(2^{T0+1}-1)}\ln 2-\frac{1}{\sigma}<0$$

for all $T0\geq 1$. In total, I conclude that

$$-2^{T0}\frac{(\ln\frac{\beta A\theta_{T0,2}^{\frac{1}{\sigma}}}{4b\psi}-\ln 2\frac{T0}{\sigma})}{2^{T0}+(\sigma-1)}-\frac{1}{\sigma}<0$$

for all $\theta_{T0,2}$. As Δ is infinitesimally small, It must be true that

$$\frac{m_i^*}{m_{i-1}^*}\Big|_{T0}>\frac{m_{i+1}^*}{m_i^*}\Big|_{T0+1}$$

for all i and $T0\geq 1$. Therefore, the span of control must fall at all existing layers when the firm adds a layer.

Next, as wage at layer i is

$$w_i(\theta)=b\psi\frac{m_i(\theta,T)}{m_{i-1}(\theta,T)},$$

wages fall at all existing layers when the firm adds a layer.

Third, as relative wage is proportional to the span of control or

$$\frac{w_{i-1}(\theta)}{w_i(\theta)}=\frac{m_i(\theta,T)}{2m_{i-1}(\theta,T)},$$

relative wages also fall at all existing layers when the firm adds a layer.

Finally, total employment increases discontinuously when the firm adds layer, as output increases discontinuously, and the span of control fall at existing layers. QED.

7.2.7 Proof of Proposition 6

Let me write out the expression of unit costs given $T+1$ layers as follows:

$$UC_T(q,b)=(2-\frac{1}{2^{T-1}})b\psi 2^{1-\frac{T}{2^T-1}}q^{\frac{1}{2^T-1}}+\frac{f_0}{q}.$$

First, taking the FOC of the above equation with respect to q results in

$$\frac{\partial UC_T(q,b)}{\partial q}=\frac{1}{2^{T-1}}b\psi 2^{1-\frac{T}{2^T-1}}q^{\frac{-(2^T-2)}{2^T-1}}-\frac{f_0}{q^2}. \tag{43}$$

There exists a unique $q_{Tm}(b)$ given T and b such that the above equation equals zero. Moreover, $\frac{\partial UC_T(q,b)}{\partial q} > 0$ if and only if $q > q_{Tm}(b)$ and vice versa. Therefore, the curve of unit costs given T and b is "U" shaped, which implies that firm productivity given T and b has an inverted "U" shape.

Next, it is straightforward to observe that

$$\lim_{q\to\infty} \frac{\partial UC_T(q,b)}{\partial q} = 0$$

given T and b. Therefore, the slope of unit costs approaches zero given T and b when output goes to infinity.

Third, the MES given T and b can be solved as follows:

$$q_{Tm}(b) = \Big[\frac{f_0}{4b\psi}\Big]^{\frac{2^T}{2^T-1}} 2^T. \tag{44}$$

The sufficiency and necessary condition for $\{q_{Tm}(b)\}_{T=1,2,3...}$ to be an increasing sequence is that

$$4f_0 > b\psi.$$

Finally, I need to derive the condition under which $\{MUC_T(b)\}_{T=1,2,3...}$ is a decreasing sequence, or

$$AVC_{T+1}(q_{T+1,m}(b),b) + \frac{f_0}{q_{T+1,m}(b)} < AVC_T(q_{T,m}(b),b) + \frac{f_0}{q_{T,m}(b)} \;\; \forall T \geq 1 \tag{45}$$

One key observation is that equation (43) implies

$$\frac{1}{2^T-1}\frac{AVC_T(q_{T,m}(b),b)}{q_{T,m}(b)} = \frac{f_0}{q_{T,m}(b)^2}. \tag{46}$$

From equation (46), I conclude that

$$AVC_T(q_{T,m}(b),b) + \frac{f_0}{q_{T,m}(b)} = (2^T-1)\frac{f_0}{q_{T,m}(b)} + \frac{f_0}{q_{T,m}(b)} = 2^T\frac{f_0}{q_{T,m}(b)}. \tag{47}$$

Substituting equation (47) into equation (45) leads to

$$MUC_T(b) > MUC_{T+1}(b),$$

if and only if

$$q_{T+1,m}(b) > 2q_{T,m}(b),$$

where $T = 1, 2, 3...$ The expression of $q_{Tm}(b)$ in equation (44) implies that

$$q_{T+1,m}(b) > 2q_{T,m}(b),$$

if and only if

$$f_0 > 4b\psi.$$

In total, if $f_0 > 4b\psi$, $\{q_{Tm}(b)\}_{T=1,2,3...}$ is an increasing sequence, and $\{MUC_T(b)\}_{T=1,2,3...}$ is a decreasing sequence. QED.

7.2.8 Proof of Proposition 7

The strategy to prove this proposition is the following. First, I assume that the incentive compatible wage defined in equation (8) satisfies the constraint indicated in equation (30) in *every* labor submarket and prove that there is a unique equilibrium with unemployment in every labor submarket. Second, I show that there is a non-empty set of parameter values within which the incentive compatible wage defined in equation (8) satisfies the constraint indicated in equation (30) in *every* labor submarket.

First, I redefine the equilibrium using three conditions. Substituting equation (10) into equation (24) leads to the homogeneous sector's employment expressed as

$$L_h = \frac{(1-\gamma)A^\sigma P^{1-\sigma}}{\gamma p_h}. \tag{48}$$

Substituting the above equation and equation (26) into equation (28) yields the following labor market clearing condition:

$$\frac{WP(\bar{\theta}, A, M) - \psi LD(\bar{\theta}, A, M)}{p_h} + \frac{(1-\gamma)A^\sigma P^{1-\sigma}}{\gamma p_h} = L. \tag{49}$$

Now, the equilibrium of the economy can be solved using three equations (i.e., equations (22), (23) and (49)). As a result, I obtain three endogenous variables: θ, A and p_h.

Values of other equilibrium variables can be solved using θ, A and p_h derived above. First, due to equation (3), I can solve the ideal price index as

$$P = \frac{1}{p_h^{\frac{1-\gamma}{\gamma}}}. \tag{50}$$

Second, the ideal price index defined in equation (4) can be reexpressed as

$$P = \Big(\int_{\theta=\bar{\theta}}^{\infty} \theta p(\theta)^{1-\sigma} M \frac{g(\theta)}{1-G(\bar{\theta})} d\theta \Big)^{\frac{1}{1-\sigma}} \equiv P_1(\bar{\theta}, A) M^{\frac{1}{1-\sigma}}. \tag{51}$$

This is because prices charged by various firms in the CES sector only depend on A and θ. Thus, the mass of firms M can be derived by using equations (50) and (51) and value of $\bar{\theta}$, A and p_h. Third, the aggregate income E can be derived by using equation (10) and value of A and P. Finally, the allocation of labor can be obtained by using equations (28) and (48) and value of A, P and p_h.

Now I show why I can use three variables (i.e., $\bar{\theta}$, A and M) to derive both the aggregate wage payment and the number of employed workers in the CES sector. In equation (9), only A and θ affect firm's optimal choices given values of exogenous parameters b and ψ. As firms endogenously choose whether or not to stay in the market, the wage payment and employment per *active* firm are functions of (A, $\bar{\theta}$) only. Therefore, I can use three variables (i.e., A, θ and M) to derive both the aggregate wage payment and the number of employed workers in the CES sector. The exact expressions of $WP(\bar{\theta}, A, M)$ and $LD(\bar{\theta}, A, M)$ can be found in the online appendix of simulation algorithm.

Next, the following claim shows the existence and uniqueness of the equilibrium in the CES sector.

Claim 3 *There exists a unique equilibrium for the CES sector characterized by a unique pair of* $(\bar{\theta}, A)$.

Proof. I have two equilibrium conditions: the ZCP condition and the FE condition. I have two endogenous variables to be pinned down: the exit cutoff $\bar{\theta}$ and the adjusted market size A. Let us think about the ZCP condition first. The goal is to establish a negative relationship between $\bar{\theta}$ and A from this condition. Suppose A increases from A_0 to $A_1 (> A_0)$ in equation (22). If the exit cutoff $\bar{\theta}$ increased from $\bar{\theta}_0$ to $\bar{\theta}_1 (\geq \bar{\theta}_0)$, the following contradiction would appear.

$$\begin{aligned} 0 &= \Pi(\bar{\theta}_1, A_1) \equiv \pi(\bar{\theta}_1, T(\bar{\theta}_1, A_1), A_1) - f \\ &\geq \pi(\bar{\theta}_1, T(\bar{\theta}_0, A_0), A_1) - f \\ &> \pi(\bar{\theta}_0, T(\bar{\theta}_0, A_0), A_0) - f = \Pi(\bar{\theta}_0, A_0) = 0. \end{aligned}$$

The first inequality comes from firm's revealed preference on the number of layers, and the second inequality is due to the fact that firm's profit function defined in equation (15) strictly increases with both θ and A. Therefore, equation (22) leads a negative relationship between $\bar{\theta}$ and A. Of course, when $\bar{\theta}$ approaches zero, A determined from equation (22) approaches infinity. And when $\bar{\theta}$ goes to infinity, A determined from equation (22) approaches zero.

Second, let me discuss the FE condition. The goal is to show that for all pairs of $(\bar{\theta}, A)$ that satisfy the ZCP condition, there is a positive relationship between these two variables determined by the FE condition. Suppose $\bar{\theta}$ decreases from $\bar{\theta}_0$ to $\bar{\theta}_1 (< \bar{\theta}_0)$ in equation (23). If the adjusted market size A increased from A to $A_1 (\geq A_0)$, the following result must be true.

$$\begin{aligned} f_e &= \int_{\bar{\theta}_1}^{\infty} \Pi(\theta, A_1) g(\theta) d\theta \\ &= \int_{\bar{\theta}_1}^{\bar{\theta}_0} \Pi(\theta, A_1) g(\theta) d\theta + \int_{\bar{\theta}_0}^{\infty} \Pi(\theta, A_1) g(\theta) d\theta \\ &> \int_{\bar{\theta}_0}^{\infty} \Pi(\theta, A_1) g(\theta) d\theta \\ &> \int_{\bar{\theta}_0}^{\infty} \Pi(\theta, A_0) g(\theta) d\theta \\ &= f_e, \end{aligned}$$

which is a contradiction. In the above derivation, I have implicitly used the ZCP condition which implies $\Pi(\theta, A_1) \geq 0$ for all $\theta \in [\bar{\theta}_1, \bar{\theta}_0]$. In total, the downward sloping ZCP curve and upward sloping FE curve intersects only once, and the intersection pins down a unique pair of $(\bar{\theta}, A)$ for the product market equilibrium.

Now, I prove the uniqueness. Suppose there were two pairs of $(\bar{\theta}, A)$ (i.e., $(\bar{\theta}_1, A_1)$ and $(\bar{\theta}_2, A_2)$) that satisfy both the ZCP condition and the FE condition. Without loss of generosity, let me assume that $\bar{\theta}_1 > \bar{\theta}_2$. Due to the property of the ZCP condition, it must be true that $A_1 < A_2$ which contradicts the positive relationship between $\bar{\theta}$ and A implied by the FE condition. Therefore, the equilibrium must be unique.

Finally, I prove the existence. For any $A \in (0, \infty)$, there exists a unique $\bar{\theta}(A)$ with $\bar{\theta}'(A) < 0$ determined by the ZCP condition. Furthermore, $\bar{\theta}(A)$ decreases continuously

in A, as the firm's profit function with the optimal number of layers increases *continuously* with θ conditional on A. Therefore, among those $(A,\ \bar{\theta}(A))$ that satisfy the ZCP condition, there must be a pair of $(\bar{\theta},\ A)$ that satisfies the FE condition. QED.

Third, the following claim shows that there is a unique p_h that clears the labor market in general.

Claim 4 *When $\frac{\sigma-1}{\sigma} \neq \gamma$ and parameter values satisfy certain conditions, there exists a unique wage p_h that clears the labor market given that the product markets are cleared.*

Proof. First, let me decompose the total wage payment of the CES sector and the number of workers employed in the CES sector into the following two parts:

$$WP(\bar{\theta}, A, M) = WPper(A, \bar{\theta}) * M$$

and

$$LD(\bar{\theta}, A, M) = LDper(A, \bar{\theta}) * M,$$

where "*per*" means per firm. Second, Substituting the above two expressions into equation (49) yields

$$\frac{WCper(A, \bar{\theta}) - \psi LDper(A, \bar{\theta})}{p_h} M + \frac{(1-\gamma)A^{\sigma}P^{1-\sigma}}{\gamma p_h} = L. \tag{52}$$

Next, substituting equation (51) into equation (3) leads to the expression of M in terms of p_h and $P_1(\bar{\theta}, A)$ as follows:

$$M = p_h^{\frac{(1-\gamma)(\sigma-1)}{\gamma}} P_1(\bar{\theta}, A)^{\sigma-1}. \tag{53}$$

Finally, substituting equations (53) and (51) into equation (49) results in the following labor market clearing condition:

$$\Big[WCper(A, \bar{\theta}) - \psi LDper(A, \bar{\theta})\Big] P_1(\bar{\theta}, A)^{\sigma-1} + \frac{(1-\gamma)A^{\sigma}}{\gamma} = p_h^{1-\frac{(1-\gamma)(\sigma-1)}{\gamma}} L. \tag{54}$$

There exists a unique p_h that satisfies the above equation, as long as $\frac{(1-\gamma)(\sigma-1)}{\gamma} \neq 1$.[44] Moreover, equilibrium p_h must satisfy the condition that

$$w_{min} \geq \psi(i) + p_h,$$

where w_{min} is the minimum wage offered in the CES sector. This puts a constraint on parameter values, which I will discuss soon. QED.

There are three effects on the labor market when the price of the homogeneous good goes up. First, as p_h is the wage offered in the homogeneous sector, labor demand of firms in the homogeneous sector goes down. Second, as p_h is the outside option for workers entering the CES sector, the number of them must go down in order to make

[44]Note that I have implicitly used the product market equilibrium conditions to derive the above equation.

the worker who chooses to enter the CES sector earn higher expected payoff.[45] These two negative effects on the labor demand are reflected by p_h that appears in the left hand side of equation (52). Finally, increasing market size due to a bigger p_h makes the aggregate income $E(= A^{\sigma}P^{1-\sigma})$ and the mass of firms M increase which pushes up the aggregate labor demand in the end. Therefore, whether or not the aggregate labor demand increases with p_h depends on whether or not the third (positive) effect dominates the first two negative effects. However, in either case, the aggregate labor demand is a *monotonic* function of p_h which assures the uniqueness of p_h that clears the labor market.

With Claim 3 and Claim 4 in hand, I only have to show that there is a non-empty set of parameter values within which the incentive compatible wage defined in equation (8) satisfies the constraint indicated in equation (30) in *every* labor submarket. In other words, I have to show that the minimum wage offered in the CES sector is weakly bigger than the wage offered in the homogeneous sector plus the disutility of exerting effort, or

$$w_{min} \geq \psi(i) + p_h.$$

First, note that labor endowment L does not affect wages and the minimum wage offered in the CES sector. This is because the solution of $(\bar{\theta}, A)$ in equilibrium does not depend on L, and wages offered by firms in the CES sector only depend on $(\bar{\theta}, A, b)$.[46] Second, equation (54) shows that p_h approaches zero when L approaches zero and $\frac{\sigma-1}{\sigma} > \gamma$, and p_h approaches zero when L goes to infinity and $\frac{\sigma-1}{\sigma} < \gamma$. Therefore, I conclude that there must exist a small enough L such that

$$w_{min} - \psi > p_h,$$

when $\frac{\sigma-1}{\sigma} > \gamma$. Similarly, there must exist a big enough L such that

$$w_{min} - \psi > p_h,$$

when $\frac{\sigma-1}{\sigma} < \gamma$.

In total, I show that with restrictions on parameter values, there must exist a unique equilibrium with unemployment in every labor submarket. The equilibrium is characterized by a unique quadruplet ($\bar{\theta}$, M, p_h, E). QED.

7.2.9 Proof of Proposition 8

This proof consists of seven parts. I prove that the exit cutoff for the quality draw increases and all firms increase the number of layers first.

I make the following notations. Suppose b decreases from $b1$ to $b2(< b1)$ due to an improvement in MT. Let $\bar{\theta}_1$ (or $\bar{\theta}_2$) be the demand threshold for exiting when $b = b1$ (or $b = b2$). Let A_1 (or A_2) be the adjusted market size when $b = b1$ (or $b = b2$).

First, I discuss how the adjusted market size A changes when b decreases by proving the following lemma.

[45] Remember that the labor demand per firm in the CES sector is independent of p_h conditional on $(A,\bar{\theta})$.

[46] Labor endowment L affects the job-acceptance-rates in various labor submarkets and accordingly the *expected* wage of entering the CES sector.

Lemma 7 *When b decrease from b_1 to b_2, the change in the adjusted market size must satisfy*

$$1 > \frac{A_2}{A_1} > \frac{b_2}{b_1}.$$

Proof. First, note that if $A_2 \geq A_1$, the exit cutoff $\bar{\theta}$ must decrease as $b2 < b1$. However, a decreasing exit cutoff plus a weakly increasing adjusted market size violate the FE condition defined in Equation (23). Thus, it must be true that $A_2 < A_1$. Second, if $\frac{A_2}{A_1} \leq \frac{b_2}{b_1}$, the profit defined as the solution to Equation (9) must decrease for all firms. Thus, the exit cutoff must increase.[47] However, the FE condition is violated again, as profit for all firms decreases, and the exit cutoff increases. In total,it must be true that

$$1 > \frac{A_2}{A_1} > \frac{b_2}{b_1}.$$

QED.

Second, I show that all firms increase the number of layers weakly. It is straightforward to observe that if $\frac{A_2}{A_1} = \frac{b_2}{b_1}$, the optimal output, employment, and the number of layers would be unchanged. As I have proven that $\frac{A_2}{A_1} > \frac{b_2}{b_1}$ in Lemma 7, all surviving firms weakly increase their number of layers.[48] Furthermore, all surviving firms increase their output as well as employment after the management technology improves.

Third, I prove that the exit cutoff increases. I use $T0 + 1 \equiv T(\bar{\theta}_1, A_1, b_1) + 1 = T(\bar{\theta}_2, A_2, b_2) + 1$ to denote the number of layers for firms on the exit cutoff and prove this result by contradiction. Suppose that the exit cutoff $\bar{\theta}$ decreased weakly after MT improves (i.e., $\bar{\theta}_2 \leq \bar{\theta}_1$). First, firms on the exit cutoff earn zero payoff due to the ZCP condition or

$$\pi(\bar{\theta}_1, T(\bar{\theta}_1, A_1, b_1), A_1, b_1) = \pi(\bar{\theta}_2, T(\bar{\theta}_2, A_2, b_2), A_2, b_2) = f,$$

as $T0 = T(\bar{\theta}_1, A_1, b_1) = T(\bar{\theta}_2, A_2, b_2)$. This leads to

$$\begin{aligned} \frac{\pi(\bar{\theta}_2, T0, A_2, b_2)}{\pi(\bar{\theta}_1, T0, A_1, b_1)} &= \left(\frac{\bar{\theta}_2}{\bar{\theta}_1}\right)^{\frac{2^{T0}}{\sigma+(2^{T0}-1)}} \left(\frac{A_2}{A_1}\right)^{\frac{2^{T0}\sigma}{\sigma+(2^{T0}-1)}} \left(\frac{b_1}{b_2}\right)^{\frac{(\sigma-1)2^{T0}}{\sigma+2^{T0}-1}\frac{(2^{T0}-1)}{2^{T0}}} \\ &\equiv X(\bar{\theta}, T0)Y(A, T0)Z(b, T0) = 1, \end{aligned}$$

where $\bar{\theta} = \frac{\bar{\theta}_2}{\bar{\theta}_1}$, $A \equiv \frac{A2}{A1} < 1$, and $b \equiv \frac{b1}{b2} > 1$. As $\bar{\theta}_2 \leq \bar{\theta}_1$,

$$Y(A, T0)Z(b, T0) \geq 1.$$

Second, For a firm whose demand draw is higher than $\bar{\theta}_1$, its profit must increase if it does not change the number of layers. This is because[49]

$$\begin{aligned} \frac{\pi(\theta, T(\theta, A_2, b_2), A_2, b_2)}{\pi(\theta, T(\theta, A_1, b_1), A_1, b_1)} &= \left(\frac{A_2}{A_1}\right)^{\frac{2^{T(\theta)}\sigma}{\sigma+(2^{T(\theta)}-1)}} \left(\frac{b_1}{b_2}\right)^{\frac{(\sigma-1)2^{T(\theta)}}{\sigma+(2^{T(\theta)}-1)}\frac{(2^{T(\theta)}-1)}{2^{T(\theta)}}} \\ &\geq \left(\frac{A_2}{A_1}\right)^{\frac{2^{T0}\sigma}{\sigma+(2^{T0}-1)}} \left(\frac{b_1}{b_2}\right)^{\frac{(\sigma-1)2^{T0}}{\sigma+(2^{T0}-1)}\frac{(2^{T0}-1)}{2^{T0}}} \geq 1, \end{aligned}$$

[47] For a rigorous proof, see Appendix 7.2.10.

[48] For a rigorous proof, see Appendix 7.2.10.

[49] Taking the log of $\frac{\pi(\theta,T,A_2,b_2)}{\pi(\theta,T,A_1,b_1)}$ leads to $B(T, A, b) \equiv \frac{2^T\sigma}{\sigma+2^T-1}\log(A) + \frac{(\sigma-1)2^T}{\sigma+2^T-1}\frac{(2^T-1)}{2^T}\log(b)$. As $B(T0, A, b) > 0$ and $\log(b) > 0$, $B(T, A, b) \geq B(T0, A, b) > 1$ for all $T \geq T0$.

where $T(\theta) \equiv T(\theta, A_1, b_1) = T(\theta, A_2, b_2)$ and $T(\theta) \geq T0$ as $\theta \geq \bar{\theta}_1$. If the firm endogenously changes the number of layers, its profit must be bigger than the profit it earns when $b = b_1$ as well due to the revealed preference argument. In total, I have

$$\pi(\theta, T(\theta, A_2, b_2), A_2, b_2) \geq \pi(\theta, T(\theta, A_1, b_1), A_1, b_1) \; \forall \; \theta \geq \bar{\theta}_1$$

for $T(\theta, A_2, b_2) = T(\theta, A_1, b_1)$ and

$$\pi(\theta, T(\theta, A_2, b_2), A_2, b_2) > \pi(\theta, T(\theta, A_1, b_1), A_1, b_1) \; \forall \; \theta > \bar{\theta}_1$$

for $T(\theta, A_2, b_2) > T(\theta, A_1, b_1)$. Third, the ZCP condition in the new equilibrium implies that firms with the quality draws between $\bar{\theta}_2$ and $\bar{\theta}_1$ earn non-negative profit. In total, the expected profit from entry would exceed the entry cost f_e if the exit cutoff decreased which violates the FE condition. Therefore, the exit cutoff must increase when b decreases.

Forth, I prove that the distribution of the number of layers moves to the right in the FOSD sense when MT improves. I make the following simplifying notations. Let $\theta_{T,2}$ be the threshold for the quality draw at which the firm increases the number of layers from $T+1$ to $T+2$. Let $Prob(t > T, b)$ be the fraction of firms that have at least $T+2$ layers when the quality of MT is b. Based on the above notations and the Pareto distribution on θ, I have

$$Prob(t > T, b) = \Big(\frac{\bar{\theta}}{\theta_{T,2}}\Big)^k.$$

Therefore, the condition for $Prob(t > T, b_2) > Prob(t > T, b_1)$ to hold is

$$\frac{\bar{\theta}_1}{\theta_{T,2}}\Big|_{b=b_1} < \frac{\bar{\theta}_2}{\theta_{T,2}}\Big|_{b=b_2},$$

where $T \geq T0$. I derive the expression for $\theta_{T,2}$ and prove the above inequality in what follows. First, conditional on (b, A), the threshold for the firm to add a layers is

$$\theta_{T,2}^{\frac{2^{T+1}}{\sigma+(2^{T+1}-1)}-\frac{2^T}{\sigma+(2^T-1)}} = \frac{b^{\frac{(\sigma-1)(2^{T+1}-1)}{\sigma+(2^{T+1}-1)}-\frac{(\sigma-1)(2^T-1)}{\sigma+(2^T-1)}}(1-\frac{\beta(2^T-1)}{2^T})\Big(\psi 2^{\frac{2^{T+2}-2-(T+1)}{2^{T+1}-1}}/\beta\Big)^{\frac{(\sigma-1)(2^{T+1}-1)}{\sigma+(2^{T+1}-1)}}}{A^{\frac{\sigma 2^{T+1}}{\sigma+(2^{T+1}-1)}-\frac{\sigma 2^T}{\sigma+(2^T-1)}}(1-\frac{\beta(2^{T+1}-1)}{2^{T+1}})\Big(\psi 2^{\frac{2^{T+1}-2-T}{2^T-1}}/\beta\Big)^{\frac{(\sigma-1)(2^T-1)}{\sigma+(2^T-1)}}}$$

Thus, the ratio of $\frac{\theta_{T,2}|_{b=b_1}}{\theta_{T,2}|_{b=b_2}}$ can be written as

$$\Big(\frac{\theta_{T,2}\big|_{b=b_1}}{\theta_{T,2}\big|_{b=b_2}}\Big)^{\frac{2^{T+1}}{\sigma+(2^{T+1}-1)}-\frac{2^T}{\sigma+(2^T-1)}} = b^{\frac{(\sigma-1)(2^{T+1}-1)}{\sigma+(2^{T+1}-1)}-\frac{(\sigma-1)(2^T-1)}{\sigma+(2^T-1)}} A^{\frac{\sigma 2^{T+1}}{\sigma+(2^{T+1}-1)}-\frac{\sigma 2^T}{\sigma+(2^T-1)}},$$

where $A \equiv \frac{A2}{A1} < 1$, and $b \equiv \frac{b1}{b2} > 1$. This expression can be simplified further to

$$\frac{\theta_{T,2}\big|_{b=b_1}}{\theta_{T,2}\big|_{b=b_2}} = (bA)^\sigma. \tag{55}$$

Second, from the expression of firm's profit function derived in Equation (15), I have

$$\frac{\bar{\theta}_1}{\bar{\theta}_2} = A^\sigma b^{(\sigma-1)(1-\frac{1}{2^{T_0}})}. \tag{56}$$

Finally, from equations (55) and (56), I conclude that

$$\frac{\frac{\bar{\theta}_2}{\theta_{T,2}|_{b=b_2}}}{\frac{\bar{\theta}_1}{\theta_{T,2}|_{b=b_1}}} = \frac{\bar{\theta}_1 b^{\sigma}}{\bar{\theta}_2 b^{(\sigma-1)(1-\frac{1}{2^{T_0}})}} \frac{\bar{\theta}_2}{\bar{\theta}_1} > 1.$$

Therefore, for all $T \geq T0$, $Prob(t > T, b_2) > Prob(t > T, b_1)$ which is the condition for the result of the FOSD to hold.

Fifth, I prove that the firm size distribution in terms of revenue moves to the right in the FOSD sense when MT improves. I make the following simplifying notations. Let $S(\bar{\theta}_i, A_i) \equiv S(\bar{\theta}_i, A_i, T(\bar{\theta}_i, A_i))_{i=1,2}$ be the revenue for firms with quality draw $\bar{\theta}_i$ when they *optimally* choose the number of layers, and $S(\bar{\theta}_i, A_i, T)$ be the revenue for firms with quality draw $\bar{\theta}_i$ when they choose to have $T+1$ number of layers. Similarly, let $q(\bar{\theta}_i, A_i) \equiv q(\bar{\theta}_i, A_i, T(\bar{\theta}_i, A_i))$ be the output for firms with quality draw $\bar{\theta}_i$ when they *optimally* choose the number of layers, and $q(\bar{\theta}_i, A_i, T)$ be the output for firms with quality draw $\bar{\theta}_i$ when they choose to have $T+1$ number of layers.

As the distribution of θ is Pareto, and the firm's revenue increases with θ, what I have to show is that for any $t > 1$,

$$S(t\bar{\theta}_2, A_2) \geq S(t\bar{\theta}_1, A_1).$$

As the distribution of the number of layers after an improvement in MT first order stochastically dominates the one before the management technology improves, I have the following two cases:

$$T(t\bar{\theta}_2, A_2) = T(t\bar{\theta}_1, A_1)$$

or

$$T(t\bar{\theta}_2, A_2) > T(t\bar{\theta}_1, A_1).$$

I discuss these two cases one by one in what follows.

In the case of $T(t\bar{\theta}_2, A_2) = T(t\bar{\theta}_1, A_1)$, if t is small enough such that $T(t\bar{\theta}_2, A_2) = T(t\bar{\theta}_1, A_1) = T_0$, then it is straightforward to see that

$$S(t\bar{\theta}_2, A_2) = S(t\bar{\theta}_1, A_1).$$

For $T(t\bar{\theta}_2, A_2) = T(t\bar{\theta}_1, A_1) = T_1 > T_0$, I have

$$S(t\bar{\theta}_2, A_2) = S(\bar{\theta}_2, A_2) V1(t, T_0, T_1) \frac{(\bar{\theta}_2 A_2^{\sigma})^{\frac{2^{T_1}}{\sigma+(2^{T_1}-1)} - \frac{2^{T_0}}{\sigma+(2^{T_0}-1)}}}{b_2^{\frac{(\sigma-1)(2^{T_1}-1)}{\sigma+(2^{T_1}-1)} - \frac{(\sigma-1)(2^{T_0}-1)}{\sigma+(2^{T_0}-1)}}}$$

and

$$S(t\bar{\theta}_1, A_1) = S(\bar{\theta}_1, A_1) V1(t, T_0, T_1) \frac{(\bar{\theta}_1 A_1^{\sigma})^{\frac{2^{T_1}}{\sigma+(2^{T_1}-1)} - \frac{2^{T_0}}{\sigma+(2^{T_0}-1)}}}{b_1^{\frac{(\sigma-1)(2^{T_1}-1)}{\sigma+(2^{T_1}-1)} - \frac{(\sigma-1)(2^{T_0}-1)}{\sigma+(2^{T_0}-1)}}},$$

where $V1(t, T_0, T_1)$ is a function of (t, T_0, T_1). As $S(\bar{\theta}_2, A_2) = S(\bar{\theta}_1, A_1) = \frac{f}{1-\frac{\beta(2^{T_0}-1)}{2^{T_0}}}$ and

$$\frac{\bar{\theta}_1}{\bar{\theta}_2} = A^{\sigma} b^{(\sigma-1)(1-\frac{1}{2^{T_0}})}$$

Based on Equation (56), I conclude that

$$\frac{S(t\bar{\theta}_2, A_2)}{S(t\bar{\theta}_1, A_1)} = \left[\left(\frac{1}{b}\right)^{(\sigma-1)(1-\frac{1}{2^{T_0}})} b^{\sigma} \right]^{(\sigma-1)\frac{2^{T_1}-2^{T_0}}{(\sigma+2^{T_1}-1)(\sigma+2^{T_0}-1)}} > 1. \tag{57}$$

In the case of $T(t\bar{\theta}_2, A_2) = T_2 > T(t\bar{\theta}_1, A_1) = T_1$, I prove $S(t\bar{\theta}_2, A_2) > S(t\bar{\theta}_1, A_1)$ using the result that when the firm optimally chooses to add a layer, output jumps up discontinuously. Note that

$$\frac{S(t\bar{\theta}_2, A_2, T_1)}{S(t\bar{\theta}_1, A_1, T_1)} \geq 1,$$

and the equality holds only when $T1 = T0$ due to Equation (57). when the firm *optimally* chooses to add layers, it must be true that

$$q(t\bar{\theta}_2, A_2, T_2) > q(t\bar{\theta}_2, A_2, T_1)$$

and

$$\begin{aligned} & S(t\bar{\theta}_2, A_2, T_2) = A_2(t\bar{\theta}_2)^{\frac{1}{\sigma}} q(t\bar{\theta}_2, A_2, T_2)^{\beta} \\ > \; & S(t\bar{\theta}_2, A_2, T_1) = A_2(t\bar{\theta}_2)^{\frac{1}{\sigma}} q(t\bar{\theta}_2, A_2, T_1)^{\beta} \\ > \; & S(t\bar{\theta}_1, A_1, T_1). \end{aligned}$$

Thus, $S(t\bar{\theta}_2, A_2)$ must be bigger than or equal to $S(t\bar{\theta}_1, A_1)$ in all possible cases. Especially, $S(t\bar{\theta}_2, A_2) = S(t\bar{\theta}_1, A_1)$ only when $T(t\bar{\theta}_2, A_2) = T(t\bar{\theta}_1, A_1) = T_0$. Therefore, the result of the FOSD for the distribution of firms' revenue follows.

Sixth, I prove the result of FOSD for the distribution of the firms' output and employment. Similar to what I have proven above, the goal is to show that for any $t > 1$,

$$q(t\bar{\theta}_2, A_2) \geq q(t\bar{\theta}_1, A_1).$$

for all $t > 1$. First, I prove that when $T(\bar{\theta}_1, A_1) = T(\bar{\theta}_2, A_2) = T0$,

$$q(\bar{\theta}_2, A_2, T_0) > q(\bar{\theta}_1, A_1, T_0).$$

To see this, note that

$$\begin{aligned} & TVC(q(\bar{\theta}_2, A_2, T_0), b_2, T_0) = \frac{\beta(2^{T_0}-1)}{2^{T_0}} S(\bar{\theta}_2, A_2, T_0) \\ = \; & TVC(q(\bar{\theta}_1, A_1, T_0), b_1, T_0) = \frac{\beta(2^{T_0}-1)}{2^{T_0}} S(\bar{\theta}_1, A_1, T_0), \end{aligned}$$

where

$$TVC(q, T, b) = (2 - \frac{1}{2^{T-1}}) b\psi 2^{1-\frac{T}{2^T-1}} q^{\frac{1}{2^T-1}},$$

$$S(\bar{\theta}_2, A_2, T_0) = S(\bar{\theta}_1, A_1, T_0) = \frac{f}{1 - \frac{\beta(2^{T0}-1)}{2^{T0}}}$$

and

$$b_1 > b_2.$$

Second, Based on the above result I derive that

$$q(t\bar{\theta}_2, A_2, T_0) = q(\bar{\theta}_2, A_2, T_0)t^{\frac{2^{T_0}-1}{\sigma+2^{T_0}-1}} > q(\bar{\theta}_1, A_1, T_0)t^{\frac{2^{T_0}-1}{\sigma+2^{T_0}-1}} = q(t\bar{\theta}_1, A_1, T_0),$$

if $T(t\bar{\theta}_2, A_2) = T(t\bar{\theta}_1, A_1) = T_0$. Third, if $T(t\bar{\theta}_2, A_2) = T(t\bar{\theta}_1, A_1) = T_1 > T_0$, I have

$$q(t\bar{\theta}_2, A_2, T_1) = q(\bar{\theta}_2, A_2, T_0)V2(t, T_0, T_1)\Big(\frac{A_2\bar{\theta}_2^{\frac{1}{\sigma}}}{b_2}\Big)^{\frac{\sigma(2^{T_1}-1)}{\sigma+(2^{T_1}-1)}-\frac{\sigma(2^{T_0}-1)}{\sigma+(2^{T_0}-1)}}$$

and

$$q(t\bar{\theta}_1, A_1, T_1) = q(\bar{\theta}_1, A_1, T_0)V2(t, T_0, T_1)\Big(\frac{A_1\bar{\theta}_1^{\frac{1}{\sigma}}}{b_1}\Big)^{\frac{\sigma(2^{T_1}-1)}{\sigma+(2^{T_1}-1)}-\frac{\sigma(2^{T_0}-1)}{\sigma+(2^{T_0}-1)}},$$

where $V2(t, T_0, T_1)$ is a function of (t, T_0, T_1). Based on equation (56), I conclude that

$$\frac{q(t\bar{\theta}_2, A_2, T_1)}{q(t\bar{\theta}_1, A_1, T_1)} = \frac{q(\bar{\theta}_2, A_2, T_0)}{q(\bar{\theta}_1, A_1, T_0)}\Big(b^{\frac{\sigma+(2^{T_0}-1)}{\sigma 2^{T_0}}}\Big)^{\frac{\sigma(2^{T_1}-1)}{\sigma+(2^{T_1}-1)}-\frac{\sigma(2^{T_0}-1)}{\sigma+(2^{T_0}-1)}} > 1,$$

as $T_1 > T_0$, $b > 1$, and $q(\bar{\theta}_2, A_2, T_0) > q(\bar{\theta}_1, A_1, T_0)$. Forth, for $T(t\bar{\theta}_2, A_2) = T_2 > T(t\bar{\theta}_1, A_1) = T_1$, I have

$$q(t\bar{\theta}_2, A_2, T_1) > q(t\bar{\theta}_1, A_1, T_1)$$

and

$$q(t\bar{\theta}_2, A_2, T_2) > q(t\bar{\theta}_2, A_2, T_1),$$

where the second inequality comes from the result that when the firm optimally chooses to add layers output jumps up discontinuously. Therefore, it must be true that

$$q(t\bar{\theta}_2, A_2, T_2) > q(t\bar{\theta}_1, A_1, T_1)$$

for $T_2 > T_1$ as well. This completes the proof for the FOSD result on the distribution of the firms' output. Finally, as firms with the same level of output (i.e., the same number of production workers) have the same employment, the result of the FOSD holds for the distribution of the firms' employment as well.

Seventh, I prove that all firms increase the span of control given the number of layers when MT improves. First, the span of control is defined as

$$SC_i(T, q(\theta, b, A, T(\theta, b, A))) = \frac{m_{i+1}(T, q(\theta, b, A, T(\theta, b, A)))}{m_i(T, q(\theta, b, A, T(\theta, b, A)))}, \tag{58}$$

where $(T-1) \geq i \geq 0$, and $q(\theta, b, A, T(\theta, b, A))$ is the number of production workers as well as output. Consider a firm with quality draw θ that does not adjust the number of layers after MT improves. This means

$$T(\theta, b_1, A_1) = T(\theta, b_2, A_2).$$

Its output and the number of production workers must increase as[50]

$$\frac{A_2}{b_2} > \frac{A_1}{b_1}.$$

The span of control calculated in equation (13) increases with the number of production workers. Therefore, *every* surviving firm increases its span of control at all layers if it does not adjusted the number of layers after MT improves. QED.

[50]For detailed proof of this result, see Appendix 7.2.10.

7.2.10 Proof of Proposition 9

I rewrite the firm's optimization problem conditional on its staying in the market (i.e., equation (9)) as

$$\begin{aligned} \max_{\{m_i\}_{i=1}^{T},T} \quad & b\Big[\frac{A}{b}\theta^{\frac{1}{\sigma}} m_T^{\frac{\sigma-1}{\sigma}} - \sum_{i=1}^{T} \psi m_i x_i\Big] \\ s.t. \quad & x_i = \frac{m_i}{m_{i-1}}, \\ & m_0 = 1. \end{aligned}$$

It is evident from the above optimization problem that b does not affect the firm's choices of employment and output *given* the number of layers and the value of $\frac{A}{b}$. Furthermore, b does not affect the optimal number of layers *conditional* on $\frac{A}{b}$, as a change in b changes profits of the hierarchy with different number of layers *proportionately* given the value of $\frac{A}{b}$. Therefore, I only need to discuss how $\frac{A}{b}$ changes when the quality of MT improves for a single firm and for a group of firms.

When MT improves for a single firm, b decreases for that firm and the adjusted market size A does not change, as each firm is "atomic". Thus, $\frac{A}{b}$ increases. An increase in $\frac{A}{b}$ has qualitatively the same effect on the firm's choices (i.e., employment, output, revenue, and the number of layers) as an increase in θ. Therefore, the firm's output, profit, and revenue increase. Furthermore, the number of layers also increases *weakly*. And the span of control increases if the number of layers does not change, since the output increases.

However, if MT improves for all other firms except for one firm, the adjusted market size A decreases as proven in Lemma 7.[51] Thus, $\frac{A}{b}$ decreases for that single firm. As a result, its output, profit, and revenue decrease. Furthermore, the number of layers also decreases *weakly*. And the span of control decreases if the number of layers does not change, since the output decreases. QED.

7.2.11 Proof of Lemma 2

I make several notations before the proof. Let A_0 be the adjusted market size for all firms in the closed economy. Denote the adjusted market size faced by non-exporters and exporters in the open economy by A_1 and $A_2(> A_1)$ respectively. As countries are symmetric, I only need to discuss what happens to domestic firms. First, note that what matters for the firm's profit is the adjusted market size A. Second, suppose that the exit cutoff stayed the same or decreased after the economy opens up to trade. This would immediately imply that $A_2 > A_1 \geq A_0$ due to the ZCP condition and the result that firms on the exit cutoff are non-exporting firm. In other words, all surviving firms are facing the bigger adjusted market size than before. However, this result together with the FE condition would imply that the exit cutoff goes up which contradicts the assumption that the exit cutoff either stayed the same or decreased. Therefore, the exit cutoff must increase after the economy opens up to trade. Third, due to this result, the adjusted market size faced by non-exporters must go down when the economy moves from autarky to trade (i.e., $A_0 > A_1$). Finally, suppose that the adjusted market size faced by exporters

[51] Again, remember that each firm is "atomic".

also went down weakly (i.e., $(A_1 <) A_2 \leq A_0$). This would imply that both exporters and non-exporters lose in the open economy. This result and the result that the exit cutoff for the quality draw increases when the economy opens up to trade would violate the FE condition. Therefore, it must be the case that $A_2 > A_0 > A_1$ in equilibrium. This means that exporters gain and non-exporters lose when the economy moves from autarky to trade. QED

7.2.12 Proof of Proposition 10

The firm's optimization problem defined in equation (9) implies that an increase (or a decrease) in A has the same effect on firm-level outcomes (i.e., revenue, employment, and output) as an increase (or a decrease) in $\theta^{\frac{1}{\sigma}}$. Proposition 5 shows that firm's revenue, employment, and output increase in θ, and Lemma 2 shows that the adjusted market size increases for exporters and decreases for non-exporters after the economy opens up to trade. Therefore, non-exporters' revenue, employment, and output go down, and exporters' revenue, employment, and output go up after the economy opens up to trade. As Proposition 1 shows that the number of layers is an increasing function of output, non-exporters de-layers and exporters increase their number of layers after the economy opens up to trade. Propositions 4 and 5 show that the firm increases the span of control when it adds a layer and decreases the span of control when it deletes a layer. Therefore, non-exporters that de-layer increase the span of control at the same time, while exporters that add a layer decrease the span of control. Finally, as wage payment in my paper is incentive-based and increases with the span of control, non-exporters increase the use of incentive-based pay when they de-layer. QED.

7.3 Supplementary Materials for Subsection 5

I discuss the details of my empirical work in this subsection. First, I describe the data sets that are used in the empirical work. Next, I explain how various variables were constructed and used. Third, I discuss the theoretical foundation for implementing the Olley-Pakes type productivity estimation in the current model. I close this subsection by showing how I implement various robustness checks for the results reported in Tables 3 and 4.

Two main data sets I use were obtained from the World Management Survey. They were originally used in Bloom and Van Reenen (2010). The first data set is a cross-sectional data set that contains roughly 5700 firms across 16 countries. Each firm was interviewed once between 2004 and 2008, and the score on each management practice (i.e., MT in hits paper) was given after the firm had been interviewed. Most interviews happened after 2006, and 2006 and 2008 are the two years that most interviews were done. The second data set is a panel data set that contains accounting information (e.g., employment, assets and sales etc.) for firms whose management scores were given in the first data set, and the time span is from 2003 to 2008. I obtain country-level data from website of either the World Bank or the Penn World Table. The Doing Business index (i.e., the distance from frontier), GDP and GDP per capita are obtained from the World Bank. The GDP deflator and the asset formation deflator are obtained from the Penn World Table. The absolute price level of countries is obtained from the Penn World Table

as well. The labor market rigidity index is included in the first data set I obtained from the World Management Survey already.

I constructed several variables for the productivity estimation. First, sales are deflated using the GDP deflator. Second, as there is no information on the use of intermediate materials, I used the Olley-Pakes approach to estimate firms' TFP. More specifically, I constructed a firm's investment by using the following equation.

$$inv_{it} = \frac{assets_{it} - 0.9 * assets_{it-1}}{the\ asset\ formation\ deflator},$$

where I assume that the value of assets depreciates by 10% annually. Finally, as the sample size is small, I grouped firms in all countries into four "big" sectors based on their SIC codes to implement the productivity estimation.

The data I used to run the regressions are observations in 2006 or 2007 mainly. The main reason is that more than 99% of the interviews were done after 2006, and firms' management quality might vary from one year to another. Thus, I chose to use accounting information after 2006 only. The second reason is that the Doing Business index (i.e., the distance from frontier) is only available from 2006. Finally, as accounting information is missing for most firms in 2008, only about thirty observations in 2008 can be used to run the regressions.

Now, I discuss the theoretical foundation for implementing the Olley-Pakes type productivity estimation in the current setup. The key to prove is that investment is a monotonically increasing function of the measured TFP conditional on the value of assets and the number of layers. I consider two types of investment in what follow. The first one is the type of investment that raises up firm's productivity. This includes R&D investment and expenditure on office automation and softwares etc. Suppose such an investment increases firm's profitability and incurs a cost. More specifically, let me assume that firm's operating profit is

$$\pi_I(A, \theta, T) = \max_I \pi_(A, \theta, T) * I - \left(\frac{I}{K}\right)^2,$$

where $\pi_(A, \theta, T)$ is defined in equation (15), and I and K are investment and value of assets respectively. Conditional on K and T, optimal investment increases with the measured TFP that is an increasing function of A and θ. This is exactly what I want to prove. The second type of investment I consider is the expenditure on capital. Namely, machines and equipments are "consumed" in the production process (i.e., depreciation). The simplest way to establish a positive relationship between this type of investment and firm's measured TFP is to assume that each unit of output "consumes" a fixed amount of capital (say, k), and all firms face the same price for capital (say, r). Under these assumptions, the firms' MC now becomes the sum of the MC derived in the paper and rk. Therefore, investment in capital is an increasing function of firm's measured productivity conditional on the number of layers and value of assets. For other production technologies such as the Cobb-Douglas production technology and the CES production technology, the logic to prove the monotonicity is similar. In total, the investment has been shown to be an increasing function of the measured TFP conditional on the number of layers and value of assets which validates the Olley-Pakes type productivity estimation in my model.

Finally, I have done several robustness checks for results reported in Tables 3 and 4. First, I re-estimated the TFP by grouping firms in each sector into *three* bins and

implementing the Olley-Pakes productivity estimation. After having obtained the new TFP estimates, I rerun the regressions specified in equations (37) and (38) and found that regression results are similar to what are reported in Tables 3 and 4. Namely, the firm-level management score positively affects firm size, ceteris paribus. Furthermore, the average management score negatively affects firm size, ceteris paribus, and the labor market rigidity index positively affects firm size, ceteris paribus. Second, I run the regressions specified in equations (37) and (38) for firms in each bin separately, as the effect of management quality on firm size might be heterogeneous across firms that have different numbers of layers. The estimation results for both groups of firms are consistent with Proposition 9 as well. Finally, it is possible that the quality of MT might differ across time for a given firm. Thus, I adjusted the data set by only keeping observations whose survey year and account year are the same. When I rerun the regressions for these remaining observations, the estimation results are again consistent with the prediction from Proposition 9. In sum, the above robustness checks confirmed the predictions established in Proposition 9.

References:

1. Atkeson, Andrew, and Ariel Burstein (2010): "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy* 118: 433-484.

2. Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta (2013): "Cross-Country Differences in Productivity: The Role of Allocation and Selection," *American Economic Review* 103: 305-334.

3. Beckmann, Martin J., (1977): "Management Production Function and the Theory of the Firm," *Journal of Economic Theory* 14: 1-18.

4. Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum (2003): "Plants and Productivity in International Trade," *American Economic Review* 93: 1268-1290.

5. Bloom, Nicholas, and J. Van Reenen (2007): "Measuring and Explaining Management Practices Across Firms and Countries," *Quarterly Journal of Economics* 122: 1341-1408.

6. Bloom, Nicholas, and J. Van Reenen (2010): "Why do management practices differ across firms and countries?" *Journal of Economic Perspectives* 24: 203-224.

7. Bloom, Nicholas, Raffaella Sadun, and J. Van Reenen (2012a): "The organization of firms across countries," *Quarterly Journal of Economics* 127: 1663-1705.

8. Bloom, Nicholas, Raffaella Sadun, and J. Van Reenen (2012b): "Management as a technology?" (Mimeo, Stanford University). Available at `http://www.stanford.edu/~nbloom/MAT.pdf`

9. Bloom, Nicholas, Benn Eifert, David McKenzie, Aprajit Mahajan, and John Roberts (2013): "Does management matter: evidence from India," *Quarterly Journal of Economics* 128: 1-51.

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

11. Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg (2012): "The Anatomy of French Production Hierarchies," NBER Working Paper 18259.

12. Calvo, Guillermo, and Stanislaw Wellisz (1978): "Supervision, loss of control, and the optimum size of the firm," *Journal of Political Economy* 86: 943-952.

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

14. Chen, Cheng (2011): "Information, Incentives and Multinational Firms," *Journal of International Economics* 85: 147-158.

15. Copeland, Brian (1989): "Efficiency Wages in a Ricardian Model of International Trade," *Journal of International Economics* 27: 221-244.

16. Davis, Donald R., and James Harrigan (2011): "Good Jobs, Bad Jobs, and Trade Liberalization," *Journal of International Economics* 84: 26-36.

17. Dixit, Avinash, and Joseph Stiglitz (1977): "Monopolistic Competition and optimum product diversity," *American Economics Review* 67: 297-308.

18. Ewing, Bradley T., and James E. Payne (1999): "The Trade-off between Supervision and Wages: Evidence of Efficiency Wages from the NLSY," *Southern Economic Journal* 66: 424-432.

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

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

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

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

23. Mookherjee Dilip (2010): "Incentives in Hierarchies" in *The Handbook of Organizational Economics* edited by Robert Gibbons and John Roberts, Princeton University Press.

24. Groshen, Erica, and Alan B. Krueger (1990): "The structure of supervision and pay in hospitals," *Industrial and Labor Relations Review* 43: 1348-1468.

25. 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.

26. Harris, John R., and Todaro, Michael P. (1970): "Migration, Unemployment and Development: A Two-Sector Analysis," *American Economics Review* 60: 126-142.

27. Hsieh, Chiang-Tai, and Pete Klenow (2009): "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics* 124: 1403-1448.

28. Hsieh, Chiang-Tai, and Pete Klenow (2012): "The Life Cycle of Plants in India and Mexico," NBER Working Paper 18133.

29. Hubbard, Thomas N. (2000): "The Demand for Monitoring Technologies: The Case of Trucking," *Quarterly Journal of Economics* 115: 533-560.

30. Hubbard, Thomas N. (2003): "Information, Decisions, and Productivity: On-Board Computers and Capacity Utilization in Trucking," *American Economic Review* 93: 1328-1353.

31. Kambourov, Gueorgui (2009): "Labour Market Regulations and the Sectoral Reallocation of Workers: the Case of Trade Reforms," *Review of Economic Studies* 76: 1321-1358.

32. Keren, Michael, and David Levhari (1979): "The Optimal Span of Control in a Pure Hierarchy," *Management Science* 25: 1162-1172.

33. Matusz, Steven J. (1986): "International Trade, the Division of Labor, and Unemployment," *International Economic Review* 37: 71-84.

34. Meagher, Kieron J. (2003): "Generalizing Incentives and Loss of Control in an Optimal Hierarchy: the Role of Information Technology," *Economics Letters* 78: 273-280.

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

36. Melitz, Marc J., and Gianmarco I. P. Ottaviano (2007), "Market Size, Trade, and Productivity," *Review of Economic Studies*, 75: 295-316.

37. Moen, Espen R. (1997): "Competitive Search Equilibrium," *Journal of Political Economy*, 105: 385-411.

38. Olley, Steven, and Ariel Pakes (1996): "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 64: 1263-1298.

39. Powell, Michael (2013): "Productivity and Credibility in Industry Equilibrium" (Mimeo, Northwestern University). Available at `http://static.squarespace.com/static/5097d207e4b06cb305096ef3/t/525dafb9e4b0924d2f46ece6/1381871545754/RelconMkts%20Oct%2015%202013.pdf`

40. Qian, Yingyi (1994): "Incentives and loss of control in an optimal hierarchy," *Review of Economic Studies* 61: 527-544.

41. Rebitzer, James (1995): "Is there a trade-off between supervision and wages? An empirical test of efficiency wage theory," *Journal of Economic Behavior and Organization* 28: 107-129.

42. Shapiro, Carl, and Joseph Stiglitz (1984): "Equilibrium unemployment as a worker discipline device," *American Economic Review* 74: 433-444.

43. Williamson, Oliver (1967): "Hierarchical Control and Optimum Firm Size," *Journal of Political Economy* 75: 123-138.

44. Yeaple, Stephen J. (2005): "A simple model of firm heterogeneity, international trade, and wages." *Journal of International Economics*, 65: 1-20.