Management Quality, Firm Organization and International Trade*

Cheng Chen†
School of Economics and Finance
The University of Hong Kong

August 18, 2014

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; 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 and have larger trade shares.

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 Mark Aguiar, George Alessandria, Pol Antràs, Costas Arkolakis, Nick Bloom, Davin Chor, Wouter Dessein, Christian Fons-Rosen, Taiji Furusawa, Stefania Garetto, Wen-Tai Hsu, Hideshi Itoh, Oleg Itskhoki, Greg Kaplan, Nobuhiro Kiyotaki, Kala Krishna, Kalina Manova, John Mclaren, Kieron Meagher, Guido Menzio, Eduardo Morales, Dilip Mookherjee, Peter Morrow, Andy Newman, Markus Poschke, Raffaella Sadun, Richard Rogerson, Thomas Sampson, Felix Tintelnot, Sharon Traiberman, John Van Reenen, Christopher Tonetti, Mu-Jeung Yang, 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.

†School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong. Homepage: `http://scholar.princeton.edu/ccfour`. E-mail: `ccfour@hku.hk`.

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 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, the organizational structure of firms with different efficiency levels and the 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 generates a selection effect that facilitates resource reallocation from less efficient and smaller firms to more efficient and bigger firms. As a result, firms are bigger, more decentralized, and more productive on average, 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 the term management technology (MT) to refer to the 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 9.1.

[3] Bloom et al. (2013) pointed 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, 2014). Most importantly, all of these are systematically related to MT. For instance, Bloom et al. (2013) argued 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 the 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 other sector which is the main focus of my analysis is a monopolistically competitive sector. It comprises a continuum of differentiated products with a constant elasticity of substitution (CES) à la Dixit and Stiglitz (1977). For simplicity I refer to the monopolistically competitive sector henceforth as the CES sector. 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]

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 a 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 selection effect of improved MT, I consider a scenario in which the quality of MT, which is common across firms, improves. This occurs when labor markets are deregulated, or better management rules are introduced into an economy. Such an improvement benefits all firms by reducing their labor costs. Furthermore, this benefit is *heterogeneous* across firms. Firms with more layers (or bigger firms) gain disproportionately more, since their average variable costs (AVCs) increase less rapidly with output. Intuitively, firms that choose to have more layers expand more aggressively after an improvement in MT. This aggressive expansion by bigger firms creates competitive pressure on smaller firms. As a result, firms with the worst demand draws are forced to leave the market, and firms with the best demand draws receive more profit and revenue. In total, the selection effect appears after an improvement in MT.[5]

The selection effect discussed above yields three implications for resource reallocation inside the CES sector. First, the resulting firm size distribution moves to the right in the first-order-stochastic-dominance (FOSD) sense after an improvement in MT. In other words, firms are bigger on average in economies with superior MT. This result is consistent with the finding from Klenow and Olken (2014) that Mexico and India whose firms have lower management scores than American firms have more small firms and fewer big firms. Moreover, average firm size is much bigger in the U.S. compared with India and Mexico. Second, 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). Finally, the aggregate productivity of firms increases as a result of an improvement in MT as well. Gains in aggregate productivity come from three sources. First, the least productive firms exit the market after an improvement in MT, which is due to the selection effect. Second, the market shares of the most productive firms increase after an improvement in MT, which is due to the selection effect as well. Third, the productivity of surviving firms increases, as improved MT reduces firm costs. In total, these three implications are the aggregate implications of an improvement in MT and are the key contributions of this paper compared to the literature. The key economic insight emphasized in this paper (i.e., the selection effect arising from an improvement in MT) is pointed out for the first

[5]In a hypothetical world, if all firms were forced to have the same number of layers due to an infinitely high cost of adding and dropping layers, the heterogeneous impact (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 heterogenous impact of an improvement in MT on firms with different demand draws.

time in the literature and deserves more attention in future research.[6]

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] 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 when they delayers. 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, trade liberalization (a trade shock) and an improvement in management quality (a management shock) complement each other due to the selection effects provided by both of them. Trade liberalization favors bigger firms, since they export. An improvement in management technology (MT) favors bigger firms as well, as they choose to adopt management hierarchies with more layers. When both shocks are present, two interesting results emerge. First, the increase in trade share (i.e., imports divided by total income) from autarky to a costly trade regime is bigger for economies with better MT. Due to the selection effect of an improvement in MT, firms are bigger and more productive on average in autarky, if the country-level management quality is higher. For a given level of reduction in the variable trade cost, this results in a disproportionate increase in the number of exporting firms relative to the total number of firms. This larger increase in the fraction of exporting firms leads to a larger trade share when the economy opens up to trade. In short, a bigger fraction of firms trade in an economy with better MT. Second, economies with superior MT benefit disproportionately from opening up to trade under certain conditions. In other words, better MT amplifies the welfare gains from trade (WGT) under certain conditions.[8] The second result is related to the first one, since the bigger increase in the trade share brings more foreign varieties to domestic consumers (the variety effect) and reduces the ideal price index of the CES goods more due to lower prices charged by foreign firms (the productivity effect). In total, the interaction between the trade shock and the management shock leads to systematic changes in aggregate trade variables.

Beyond the 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 at existing layers when the firm adds a layer. This reduction occurs because the addition of a new layer reduces the span of control at existing layers. This result shows that employees might lose from firm's expansion. Second, when wages increase,

[6]Powell (2013) investigated how contract enforcement affects the *dispersion* of the distribution of firm productivity in a perfectly competitive product market. One key result from that 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) focused on the second-order moment of the distribution of firm productivity, while my paper focuses on the first-order moment of it.

[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 my model, WGT are not guaranteed. Whether or not there are WGT depends on parameters values such as the elasticity of substitution and the quality of MT.

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 of firm expansion on workers' wages . 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" organization 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.

The current paper also quantitatively evaluates how an improvement in MT affects firm size, aggregate productivity and welfare. In order to implement counterfactual experiments, I calibrate the model to match six moments obtained from the data of the U.S. economic census. Then, I implement two counterfactual experiments by improving the management quality for all firms in the economy. Calibration results show that an improvement in MT has quantitatively important effects on aggregate productivity, firm size and welfare. More specifically, a 22% improvement in MT results in a 22.2% increase in the weighted average of firm productivity, a 72.2% increase in average employment, and a 47.1% increase in welfare. Furthermore, its impact on the WGT is quantitatively sizable as well. The WGT increase by about 1.21% after the improvement in MT. In short, the calibrated model is able to generate quantitatively important effects of an improvement in MT on aggregate-level economic outcomes.

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 manage to derive the optimal number of layers for the firm by characterizing the firm's cost functions. This is for the first time in the research of incentive-based hierarchies that such a difficult problem has been solved. 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 in my model, as each firm faces a downward-sloping demand curve. The determination of optimal firm size is a central theoretical obstacle for the research on incentive-based hierarchies. This paper manages to solve this problem by considering the firm's problem in a monopolistically competitive market environment.[10] 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 weighted average of firm productivity, 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 eco-

[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 (1978) pointed out that firm size is undetermined in the standard model of incentive-based hierarchies. Firm size in Qian (1994) is exogenously given, as the firm is assumed to have a fixed amount of capital and a Leontief technology to produce. If the firm is allowed to choose the amount of capital optimally, firm size goes to infinite, as pointed out by Meagher (2003).

nomic 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 that is closely related to mine is Caliendo and Rossi-Hansberg (2012). There are three major differences between these two papers. First, the moral hazard problem which is absent in Caliendo and Rossi-Hansberg (2012) is a 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, the focus of my paper is the selection effect of better MT in the *closed* economy, while Caliendo and Rossi-Hansberg (2012) focused on how trade liberalization affects firm organization and productivity in the *open* economy. Finally, this paper yields some important micro-level predictions that Caliendo and Rossi-Hansberg (2012) did not have. For example, the result that firms increase the number of layers when MT improves is unique to my model, since improvements in the ICT would imply flattened firm hierarchies (i.e., delayering). This difference is important, since the number of layers has been shown to be an important factor determining firm performance and workers' wages (Caliendo, Monte, and Rossi-Hansberg (2012)).

This article contributes to the recent macro-development literature that discusses cross-country differences in the firm size distribution and resource misallocation (e.g., Hsieh and Klenow (2009, 2012) and Hsieh and Olken (2014)) in two ways. First, this paper provides a new explanation for why firms in developing countries are smaller and less productive on average compared to those in developed countries. As pointed out by Hsieh and Olken (2014), some prevailing arguments that are used to explain cross-country differences in the firm size distribution face empirical challenges. For instance, size-dependent policies which are commonly used in many developing countries would imply bunching of firms around certain thresholds on firm size. However, this type of bunching does not seem to be quantitatively important in the data, at least for India, Mexico and Indonesia (Hsieh and Olken (2014)). Another popular view is that financial constraints hurt *smaller* firms more than bigger firms in developing countries. This argument would imply that developing countries have "missing middle" in the size distribution of firms. That is, developing countries have both more small firms *and* more big firms compared with developed countries. Moreover, there are fewer medium-sized firms in developing countries compared with developed countries as well (i.e., a bimodal distribution of firm size). Unfortunately, this feature does not seem to exist in the data as well (Hsieh and Olken (2014)). On the contrary, the theory proposed in this paper does not have the above counterfactual predictions. Second, the key economic insight of this paper that bigger and more efficient firms are constrained more in developing countries (i.e., countries with worse MT) is consistent with conjectures and suggestive evidence provided by Bloom et al. (2013) and Hsieh and Olken (2014). In summary, this essay points out a new channel through which differences in quality of MT affect the firm size distribution and resource allocation. Furthermore, aggregate predictions of the theory are consistent with the suggestive evidence.

This paper 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)). The current paper complements this literature by showing how the organizational structure of the firm and the quality of MT affect the responses of

firms and economies to trade liberalization. In particular, the WGT are impacted by the quality of MT which is affected by institutional quality.

The current paper is also related to the literature that applies efficiency wage models into the international context. Early contributions include Copeland (1989) and Matusz (1996). More recently, Chen (2011) investigated how the consideration of the efficiency wage affects multinational firms' organizational choices (i.e, FDI or outsourcing). Davis and Harrigan (2011) used the standard efficiency wage model (i.e., Shapiro and Stiglitz (1984)) to discuss how trade liberalization affects wages of various jobs (i.e., jobs in non-exporting firms and exporting firms etc.) differently. The key departure of this paper from the existing literature is the endogenous determination of the monitoring intensity (i.e., management quality). Firms endogenously determine their monitoring intensities, since they make organizational choices. The endogenous formation of internal firm organization is the key to generating the results on the changes in aggregate trade variables.

The remainder of the paper is organized as follows. Section two solves the individual firm's optimization problem. Section three solves the problem of resource allocation in general equilibrium. Section four investigates how differences in MT across economies affect various aggregate economic outcomes. Section five extends the baseline model into an international context à la Melitz (2003) to investigate how the internal organization of firms responds to bilateral trade liberalization. Section six studies how the quality of MT affects the WGT as well as the trade share by treating the number of layers as a continuous variable. Section seven calibrates the model to quantitatively evaluate how an improvement in MT affects various aggregate economic outcomes including aggregate productivity and the WGT. Section eight concludes.

2 Partial Equilibrium Analysis

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 in the next section.

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 variety produced by a firm, 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

$$\mathbf{C} \equiv \Big(\frac{C_c}{\gamma}\Big)^{\gamma} \Big(\frac{C_h}{1-\gamma}\Big)^{1-\gamma}, \tag{3}$$

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

$$P^{\gamma} p_h^{1-\gamma} \equiv 1, \tag{4}$$

where

$$P = \Big(\int_{\Omega} \theta p(\theta)^{1-\sigma} M\mu(\theta) d\theta \Big)^{\frac{1}{1-\sigma}} \tag{5}$$

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 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 the 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 obtained 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 free entry of firms in equilibrium.[11] 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 f_2 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 their direct subordinates in order to incentivize them to work.[12]

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.[13] 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

$$q = \int_0^{m_T} a(j)dj, \tag{6}$$

where m_T is the measure 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, because there is a continuum of workers. 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.[14] Layer T is the lowest layer in the hierarchy which is occupied by production workers. In short, 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 cannot 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

$$w_i - \psi \geq (1 - p_i)w_i, \tag{7}$$

[11] Essentially, I assume that there is another sector which is not explicitly modelled here. The entrepreneur can enter and receive the payoff of f_2 by working in that sector.

[12] 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 jobs 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 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' jobs in order to distill the key economic impact of an improvement in MT on economic outcomes.

[13] As shown in Appendix 9.2, it is straightforward to generalize the analysis to allow effort to be a continuous variable.

[14] A smaller i denotes a higher layer in the firm's hierarchy, and the entrepreneur is at layer zero.

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 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 $p(b, x_i)$:

$$p(b, x_i) = \frac{1}{bx_i(\theta)}, \tag{8}$$

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' effort inputs.[15] Parameter b reflects the *inefficiency* of MT. More specifically, the worse the MT is, the bigger the value of b.

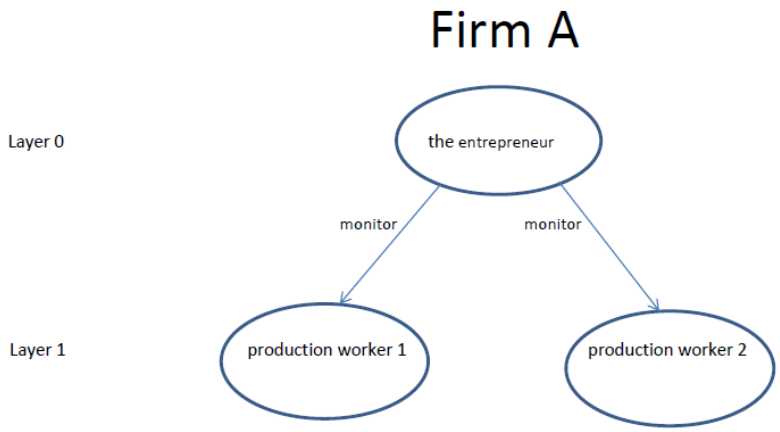
A firm may want to hire non-production workers, since it wants to economize on the cost to incentivize workers. I use Figures 1 and 2 to clarify the economic intuition behind this choice.[16] Consider firm A that receives a low demand draw θ_A and wants to produce two units of goods as illustrated in Figure 1. The span of control of the entrepreneur is small for this firm, which implies a low incentive-compatible wage paid to production workers. Thus, it is optimal to have production workers only, since non-production workers do not produce output. Next, consider firm B that receives a high demand draw θ_B 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 to have *multiple* layers of non-production workers when the output level is high. In total, the trade-off between lower wages paid to

[15] As what I will show, the entrepreneur allocates the monitoring intensities 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 derivations are available upon request.

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

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

Figure 1: A Firm with Two Layers



existing workers and extra wages paid to workers employed at the new layer determines 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 intensities evenly across workers at a given layer. Intuitively, 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 9.3.1.

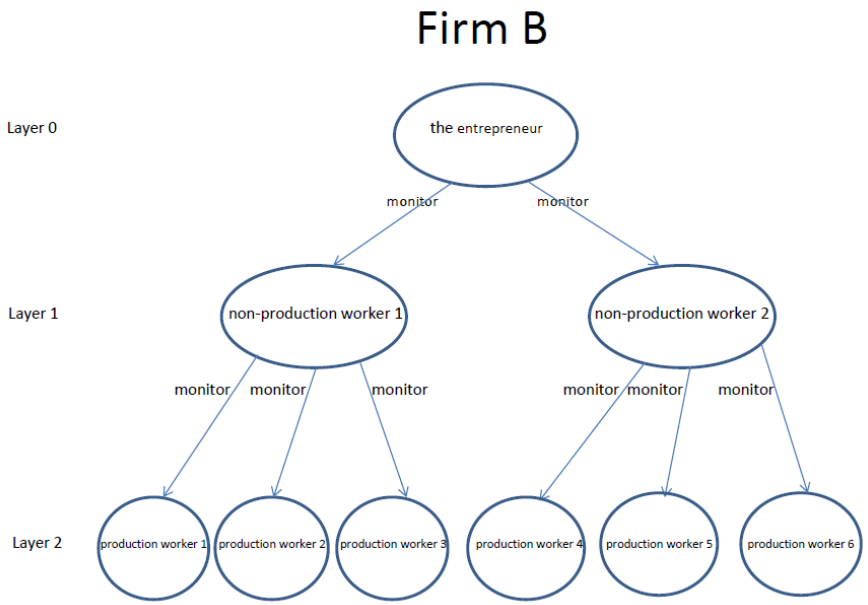
The entrepreneur faces the same incentive problem as her employees. She sits at the top of the hierarchy and is monitored by nobody. However, the entrepreneur of any firm that chooses to *stay* in the market is incentivized to exert effort in equilibrium (i.e., monitor her subordinates). First, staying in the market and shirking result in zero output 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 (8) into inequality (7), I derive the minimum incentive-compatible wage for layer i as follows:

$$w_i(\theta) = \frac{\psi}{p_i(b, x_i(\theta))} = \psi b x_i. \tag{9}$$

The key feature of the above equation is that the minimum incentive compatible wage $w_i(\theta)$ is negatively related to the supervision intensity. This relationship finds support

Figure 2: A Firm with Three Layers



in the data; see, for example, Rebitzer (1995) and Groshen and Krueger (1990).[18] Based on equations (1), (6), (9) 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

$$\begin{aligned} \max_{\{m_i\}_{i=1}^{T},T} \quad & A\theta^{\frac{1}{\sigma}} m_T^{\frac{\sigma-1}{\sigma}} - \sum_{i=1}^{T} b\psi m_i x_i \\ s.t. \quad & x_i = \frac{m_i}{m_{i-1}}, \\ & m_0 = 1. \end{aligned} \tag{10}$$

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:

$$A \equiv \left(\frac{\gamma E}{P^{1-\sigma}}\right)^{1/\sigma}, \tag{11}$$

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

[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.

imply

$$w_T m_T = 2w_{T-1}m_{T-1} = ... = 2^{T-1}w_1 m_1, \tag{12}$$

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

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

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

$$x_i(q,T) = \frac{m_{i+1}(q,T)}{m_i(q,T)} = 2\Big(\frac{q}{2^T}\Big)^{\frac{2^{T-(i+1)}}{2^T-1}}, \tag{14}$$

which increases with q given the number of layers. Equation (13) shows that employment at each layer increases with output given the number of layers. Moreover, equation (14) 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

$$q(\theta,T) = m_T(\theta,T) = \left[\frac{A\beta\theta^{\frac{1}{\sigma}}}{b\psi 2^{2-\frac{T}{2^T-1}}}\right]^{\frac{\sigma(2^T-1)}{\sigma+(2^T-1)}}, \tag{15}$$

which is the firm's output level as well as the number of production workers. Substituting equations (13) and (15) into equation (10) leads to the firm's operating profit (i.e., profit before paying the fixed cost) and revenue as

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

and

$$S(\theta,T) = (A\theta^{\frac{1}{\sigma}})^{\frac{2^T\sigma}{\sigma+(2^T-1)}} \left(\frac{\beta/b\psi}{\Big(2^{\frac{2^{T+1}-2-T}{2^T-1}}\Big)}\right)^{\frac{(\sigma-1)(2^T-1)}{\sigma+(2^T-1)}}, \tag{17}$$

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 to solve 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 a management hierarchy with $T+1$ layers. Based on equations (13) and (14), $TVC_T(q,b)$ is derived as

$$TVC_T(q,b) = \sum_{i=1}^{T} m_i(q,T) w_i(q,T) = \sum_{i=1}^{T} b\psi \frac{m_i^2(q,T)}{m_{i-1}(q,T)} = (2 - \frac{1}{2^{T-1}}) b\psi 2^{1-\frac{T}{2^T-1}} q^{\frac{2^T}{2^T-1}}. \tag{18}$$

Variable cost given the number of layers increases with output. Better MT, which is denoted by a smaller value of b, pushes down the variable cost given any number of layers proportionately.

With the firm's cost functions given different numbers of layers in hand, I can characterize 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 9.3.2.

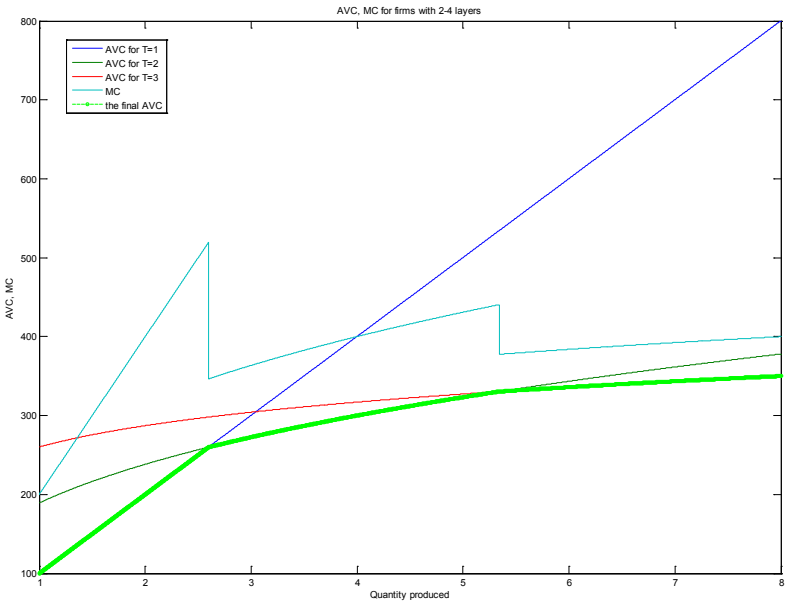
Figure 3 illustrates the AVC curve and the MC curve. The AVC curve denoted by the bold green curve is the lower envelope of all AVC curves given different numbers of layers. The MC curve does not increase with output monotonically. The span of control increases at all layers when the firm increases output 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. 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

[19]For firms that have one layer (i.e., self-employed entrepreneurs), management hierarchies are not needed. As this paper focuses on management hierarchies, I do not consider these firms in the paper.

Figure 3: Average Variable Cost and Marginal Cost



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 incentives 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.*

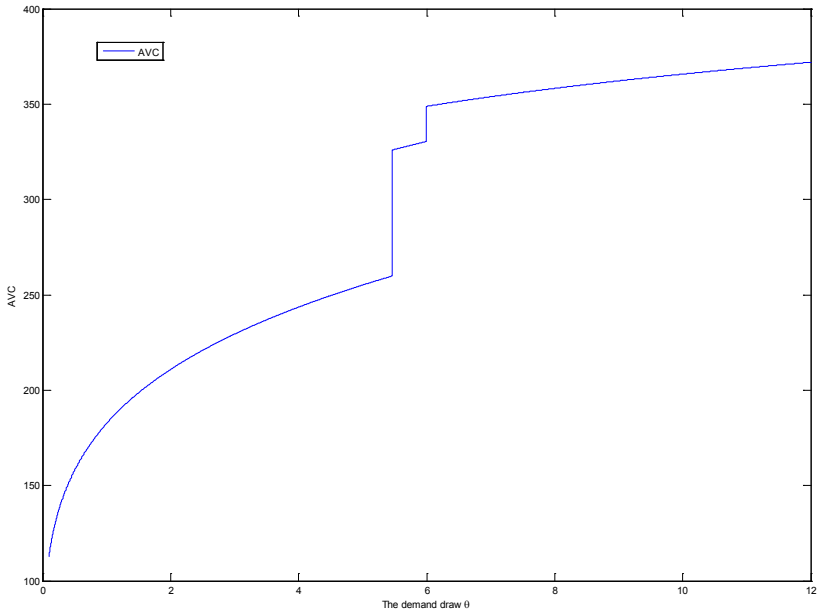
Proof. See Appendix 9.3.3.

Proposition 2 and the distinctive feature of the AVC curve discussed above are the keys to understanding the selection effect of an improvement in MT. Improved MT reduces labor costs 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 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 are bigger are found to have more layers as well. All these evidence supports the 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 a change in the firm's demand draw. These results are useful, since I am going to analyze the firm size distribution in the next section. Proposition 3 summarizes the results.

Figure 4: Average Variable Cost and the Demand Draw



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 9.3.4.

The strategy of a firm to grow depends crucially on whether or not the production is reorganized. When the firm grows 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]

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, the relative wage, defined as the ratio of the supervisor's wage to his direct subordinate's wage, behaves in a way consistent with the findings of Caliendo, Monte, and Rossi-Hansberg (2012).

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

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 9.3.5.

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. 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.*

Proof. See Appendix 9.3.6.

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 after an addition of a layer, since the span of control is reduced.

[21] As in 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.

Consequently, it is an efficient way to economize on labor cost by reducing their wages disproportionately more.

The results of Proposition 5 are 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 in this subsection. Following Caliendo and Rossi-Hansberg (2012), I use the inverse of unit costs (i.e., output divided total costs) to measure firm productivity and analyze how the unit costs vary with output as well as the number of layers.

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

$$UC_T(q,b) \equiv \frac{TVC_T(q,b)+f_0}{q} = AVC_T(q,b) + AFC(q), \tag{19}$$

where f_0 is the fixed production cost, and $AFC(q)$ is the average fixed cost. The unit costs given the inefficiency of MT are defined as

$$UC(q,b) \equiv \frac{TVC(q,b)+f_0}{q} = AVC(q,b) + AFC(q). \tag{20}$$

As the fixed production cost does not affect the choice of the number of layers, equation (20) 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 is "U"-shaped. 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 threshold. 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 curves of the 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 unit costs given the 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

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

And the MUC given b and T is defined as

$$MUC_T(b) \equiv UC_T(q_{Tm}(b), b). \tag{22}$$

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 a lower MUC.

Assumption 1 $f_0 > 4b\psi$.

Under this assumption, firm productivity has an increasing overall trend with respect to output. On the contrary, firm productivity has an decreasing overall trend with respect to output, if Assumption 1 is violated.[22] Economically, Assumption 1 requires that MT is efficient enough (i.e., b is small enough). I assume that Assumption 1 holds in what follows, since the estimated parameters from a calibrated model presented presented in Section 7 satisfy this constraint.

Now, I characterize 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 9.3.7.

The inefficiency of MT is the key to understanding this proposition. Improved MT makes the MES increase and the MUC decrease given the 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 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 *Given the number of layers, the curve of firm productivity is inversely "U"-shaped. Under Assumption 1, the maximum value of firm productivity given the number of layers increases with the number of layers.*

[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) found that the positive correlation between firm productivity and size is much stronger in developed countries such as the U.S. and Germany compared with developing countries such as Hungary and Slovenia. In Romania, the covariance between firm productivity and size is even negative. Although there are no management data for firms in Hungary, Slovenia and Romania, firms in Poland which is also a central European country have lower management scores compared with firms in the U.S. and Germany. Therefore, the finding in Bartelsman, Haltiwanger, and Scarpetta (2013) supports my productivity measure.

Figure 5: Firm Size and Productivity

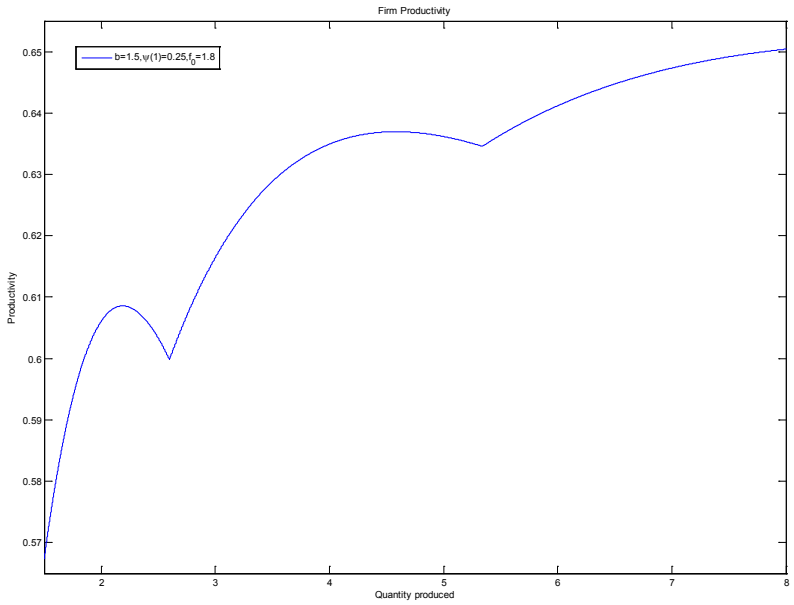


Figure 5 illustrates how firm productivity varies with output. The overall trend is that firm productivity increases with output. This implies that small firms are less productive on average. Aggregate productivity increases if the smallest firms exit the market after some shocks. This is one impact of improved MT on aggregate productivity which I will analyze in the next section.

3 General Equilibrium Analysis

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 sectors and various labor submarkets in equilibrium, since they can freely move across sectors and firms.

3.1 Product Market Equilibrium

There are two equilibrium conditions for the CES sector: the zero cutoff *payoff* (ZCP) condition and the free entry (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 active firms in equilibrium (i.e., M). First, the ZCP condition that firms with the quality draw $\bar{\theta}$ earn

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

zero payoff can be written as

$$\Pi(\bar{\theta}, A) = 0, \tag{23}$$

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 obtained from entering the CES sector equals the outside option of entrepreneurs, or

$$\int_{\bar{\theta}}^{\infty} \Pi(\theta, A) g(\theta) d\theta = f_e, \tag{24}$$

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

$$p_h L_h = (1 - \gamma) E, \tag{25}$$

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.

3.2 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 homogeneous sector, which is the outside option of workers entering the CES sector. In total, I have

$$\frac{m_i(\theta)}{Q(\theta, i)}(w_i(\theta) - \psi) = \frac{m_{i'}(\theta')}{Q(\theta', i')}(w_{i'}(\theta') - \psi) = p_h \quad \forall\, (i, i') \,\forall\, (\theta, \theta'), \tag{26}$$

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 obtained 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 worker's indifference condition in equation (26), or

$$\begin{aligned} L_c &= \int_{\theta=\bar{\theta}}^{\infty} \sum_{i=1}^{T(\theta,A)} Q(\theta, i) \frac{Mg(\theta)}{1 - G(\bar{\theta})} d\theta \\ &= \frac{WP(\bar{\theta}, A, M) - \psi LD(\bar{\theta}, A, M)}{p_h}, \end{aligned} \tag{27}$$

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

$$LD = \int_{\theta=\bar{\theta}}^{\infty} \sum_{i=1}^{T(\theta,A)} m(\theta, i) \frac{Mg(\theta)}{1 - G(\bar{\theta})} d\theta. \tag{28}$$

Equation (27) has an intuitive explanation. It says that the *total* expected payoff of workers entering the CES sector (i.e., $p_h L_c$ 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

$$L_h = L - L_c. \tag{29}$$

Equations (27) and (29) are two labor market equilibrium conditions that are used to determine the allocation of labor between sectors.

3.3 Equilibrium and Unemployment

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

$$E = \int_{\bar{\theta}}^{\infty} LC(\theta) \frac{Mg(\theta)}{1 - G(\bar{\theta})} d\theta + p_h L_h + \Big[f_0 + f_1 \big(\frac{\bar{\theta}}{\theta_{min}} \big)^k \Big] M + \Big[\psi + f_2 \big(\frac{\bar{\theta}}{\theta_{min}} \big)^k \Big] M, \tag{30}$$

where $LC(\theta)$ is the total wage payment of firms with demand draw θ. The third part of the right hand side (RHS) of equation (30) is the demand for the final composite good by firms, and the last part of the RHS of equation (30) is the consumption of active

[24] For further discussion of the labor market equilibrium in the CES sector, see Appendix 9.3.8.

entrepreneurs who earn profit in equilibrium.[25] Total income of the economy equals total expenditure which includes two parts: demand from workers and demand from firms. Note that only workers and *active* entrepreneurs demand goods in equilibrium. Entrepreneurs who choose not to enter the CES sector receive their outside option without consuming goods; 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 (23), (24), (25), (27), (29) and (30)). Obviously, one equilibrium condition is redundant due to Walras' law, and I normalize the price of the final composite good to one.

One implicit assumption for the existence of the equilibrium is that the probability of being employed implied by equation (27) is smaller than or equal to one in *every* labor submarket in equilibrium. In other words, wages offered in the CES sector must satisfy

$$w_i(\theta) - \psi \geq p_h \ \ \forall\, (i, \theta), \tag{31}$$

where $w_i(\theta)$ is determined in equation (9). The above inequality would be violated if ψ were zero. Firms do not need to pay incentive-compatible wages to workers, if exerting effort does not generate any cost to them. At the same time, there is no unemployment in *all* labor submarkets, and every worker in the CES sector receives the same wage. I don't consider this case, since information friction and MT do 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 (9) satisfies the constraint specified in equation (31) in *every* labor submarket. There are three reasons why I want to investigate this type of equilibrium. First, the model yields clean insights and testable implications in this case. Second, the testable implications at the micro-level (e.g., wages, relative wages, and the optimal number of layers etc.) have been shown to be consistent with the evidence presented in Caliendo, Monte, and Rossi-Hansberg (2012). Finally, as shown below, the model's predictions at the aggregate level (i.e., effects of improved MT on the firm size distribution and firm organization) 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]

Proposition 7 *When $\frac{\sigma-1}{\sigma} > \gamma$, there exists a unique equilibrium with unemployment in every labor submarket, if the labor endowment (i.e., L) is small enough; When $\frac{\sigma-1}{\sigma} < \gamma$,*

[25] The equilibrium condition stated in equation (25) has used the FE condition described above. Ex post profit per active firm must compensate both the cost of exerting 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 of exerting effort is ψM. Therefore, profit per active firm that 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. In this case, firms that would offer lower incentive-compatible 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 Appendix 9.6 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.

there exists a unique equilibrium with unemployment in every labor submarket if the labor endowment is big enough.

Proof. See Appendix 9.3.8.

The labor endowment affects the outside option of workers through two channels. First, a bigger labor endowment reduces workers' wage in the homogeneous sector as a result of the supply-side effect. Second, a bigger labor endowment increases total income of the economy, which leads to a larger demand for labor and increases the wage in the homogeneous sector. The relative strength of these two effects depends on parameter values. When $\frac{\sigma-1}{\sigma} > \gamma$, the demand-side effect dominates. Thus, a sufficiently small labor endowment insures that the workers' outside option is small enough in equilibrium, which validates the existence of a unique equilibrium with unemployment in every labor submarket and vice versa.[27]

Admittedly, the condition assuring the existence of a unique equilibrium with unemployment in every labor submarket involves endogenous variables.[28] This is because both wages offered in the CES sector and the wage offered in the homogeneous sector cannot be solved analytically. However, when I treat the number of layers as a continuous variable à la Keren and Levhari (1979) and Qian (1994), the condition can be stated using exogenous parameters only. Readers are referred to Appendix 9.4 for more details.

Aggregate labor demand that takes into account the product-market-clearing conditions either increases or decreases in p_h. There are two countervailing effects on the aggregate labor demand, when the workers' outside option increases. On the one hand, the homogeneous sector demands less labor when p_h goes up, and the number of job applicants that equalizes the expected payoff obtained from entering the two sectors goes down due to the higher outside option. On the other hand, the higher expected wage increases the total income of the economy, which makes both sectors demand more labor. The relative strength of these two offsetting effects depends on parameter values. When $\frac{\sigma-1}{\sigma} < \gamma$, the first effect dominates and vice versa. $\frac{\sigma-1}{\sigma} = \gamma$, the aggregate labor demand does not respond to the change in 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 the 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 the 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.[29] 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

[27]The wage determination in the CES sector features that buyers (i.e., firms) have all the bargaining power when the wage contracts are offered. This implies that changes in labor endowment do not affect incentive-compatible wages, as long as there is unemployment in the labor submarkets.

[28]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.

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

rates) across different labor submarkets are used to equalize the expected payoff obtained from entering different labor submarkets. Essentially, the role of unemployment rate in my model is the same as in Harris and Todaro (1970) and is similar to the role played by the labor market tightness in the literature on competitive search (e.g., Moen (1997) etc.).[30]

4 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, 2012) showed that China and India whose firms receive low management scores have more firms of a small size and fewer efficient firms with large market shares than the U.S. Finally, firm organization differs across countries due to differences in MT and affects firm size and performance as well. More specifically, 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 firms' expansion (e.g., Bloom, Sadun, and Van Reenen (2012a), 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 characteristics discussed above.

4.1 Selection Effect of Better Management Technology

I consider a scenario in which the MT that is common across all firms improves. Such an improvement is equivalent to a decrease in b in the model, since 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, since they have more layers. More specifically, an improvement in MT benefits all firms since 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. As a result, firms with the worst demand draws are forced to leave the market; firms whose demand draws are in the middle receive shrinking revenue and profit; and firms with the best demand draws expand. In other words, an improvement in MT facilitates inter-firm resource allocation

[30]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 future 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.

through benefitting bigger firms more, which is exactly what Bloom et al. (2013) argued in their paper.

Endogenous selection of a management hierarchy with a specific number of layers is the key to understanding the selection effect of an improvement in MT. In a hypothetical world, if all firms were forced to have the same number of layers, the uneven effect would disappear. This is because all firms would have the same AVC function in such a world. As a result, the exit cutoff for the demand draw would be unaffected by an improvement in MT. Furthermore, firms' revenue and profit would be unchanged as well. In short, the pro-competitive effect (i.e., the selection effect) 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) found that Indian firms are unwilling to decentralize their production processes (i.e., constrained span of control), because it is hard to catch and punish misbehaving employees in India. Furthermore, they argued that poor monitoring and weak enforcement of laws are reasons for why Indian firms can't catch and punish misbehaving workers easily. Finally, they argued that low level of decentralization is one reason for why Indian firms are small on average. My model gives economic reasons rationalizing these findings. First, when firms are able to monitor their employees more easily, the span of control increases. 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.[31] 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 shown below. In total, firms are less decentralized in economies with poor MT, and the low level of decentralization is one reason for why firms are small in these economies.

The predictions on the internal structure of firms are unique predictions of my model. In knowledge-based hierarchy models (e.g., Garicano (2000)), an improvement in ICT flattens firms' hierarchies. That is, firms de-layer when ICT improves (See Garicano and Van Zandt (2012) for details). The intuition is that the impact of an improvement in ICT is heterogeneous across layers *conditional* on output. When communication becomes more efficient, firms increase the knowledge learned by workers at upper layers and reduce the knowledge learned by workers at lower layers. Furthermore, the number of layers is reduced as well. When the costs of learning knowledge become cheaper, firms increase the knowledge learned by workers at all layers and decrease the number of layers. However, an improvement in MT does not have such a heterogeneous effect across layers in this paper, since monitoring is done layer by layer. In other words, two non-adjacent layers do not interact with each other in the management hierarchy considered in this paper. Therefore, better MT affects different layers evenly. As a result, firm size (i.e., output or employment) is a sufficient statistic to determine the optimal number of layers.

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

$$G(\theta) = 1 - \left(\frac{\theta_{min}}{\theta}\right)^k, \tag{32}$$

[31] Note that output and employment go up for *all* firms. However, revenue and operating profit fall for small firms after MT improves.

where $G(\theta)$ is the CDF of the demand draw θ. 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.

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 First-Order-Stochastic-Dominance (FOSD) sense.*

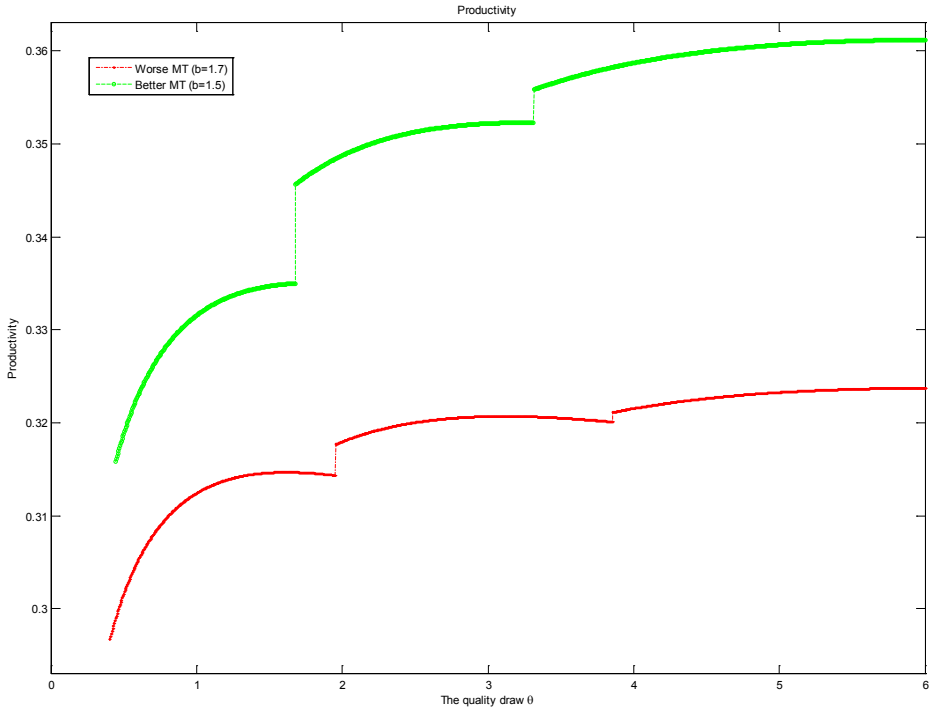
Proof: See Appendix 9.3.9.

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 à la Keren and Levhari (1979) and Qian (1994).[32] In total, the aggregate-level predictions (i.e., changes in the exit cutoff, the firm size distribution, and the number of layers) derived in the case of a continuous number of layers are qualitatively the same as those derived in the case of a 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 implies that there are fewer small firms in economies with superior MT. Furthermore, firms with better demand draws have bigger sales, and average firm size is bigger in such economies as well. These theoretical predictions show the key role played an improvement in MT in determining resource allocation. 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 efficient firms don't have large enough market shares in such economies. Finally, these theoretical predictions are consistent with several key findings and conjectures of a number of recent papers. For instance, Hsieh and Klenow (2009) showed that India and China are worse at getting efficient firms to obtain big market shares compared with the U.S. Bloom, Sadun, and Van Reenen (2012b) found that a given amount of increase in the management score has a bigger positive impact on firm size in the U.S. than in other countries. As management scores are highly correlated with firm productivity, this implies that more productive firms gain more in economies with better MT such as in the U.S. This finding is exactly the key prediction of my model. Next, Hsieh and Olken (2014) argued that it is the big firms that are constrained more

[32]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 Appendix 9.4 for details.

Figure 6: Gains in Aggregate Productivity



in developing countries, not the small firms. This conjecture is also consistent with the key result of this paper. In short, the key results of the model seem to square well with existing evidence.

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. The weighted average of firm productivity is defined as the sum of the product of firm productivity and its market share across all firms.[33] Gains in the weighted average of firm productivity come from three sources. First, firms with the worst demand draws which are less productive on average exit the market after an improvement in MT (i.e., the between-firm effect). Second, market shares of more productive firms increase, because improved MT favors more productive firms. This makes the weighted average of firm productivity increase as well (i.e., the between-firm effect). Finally, the productivity of all surviving firms increases, as improved MT reduces firms' costs (i.e., the within-firm effect). 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 worker's welfare. As workers can freely move between two sectors, the expected payoff obtained from entering the CES sector must equal the wage offered in the homogeneous sector. Therefore, the worker's expected payoff obtained from entering the CES sector is

[33]The market share is either the firm's share in the sector's total output or its share in the sector's total sales.

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 the 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 does not necessarily increase welfare. It turns out that the factor governing the direction of the change in welfare is the elasticity of substitution between products in the CES sector, since 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 wages and higher probabilities 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., the increase in average 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, the bigger increase in average employment 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. This increase reduces the risk of being unemployed for workers in the CES sector. In summary, the increase in the average probability of being employed dominates wage loss, when MT improves and the elasticity of substitution is high. As a result, the worker's expected payoff obtained from entering the CES sector increases.

The opposite story happens when the elasticity of substitution is low. 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, workers obtain a lower expected payoff by 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 the MT improves, the ideal price index for the CES sector 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.

The above discussions on welfare are not analytical results, although 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. Importantly, when the number of layers is treated as a continuous variable, welfare increases after an improvement in MT if and only if $\frac{\sigma-1}{\sigma} > \gamma$. This qualitative result is shown in Appendix 9.4.

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

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

	Welfare	Ave(wage)	Ave(ur)	M	E
b=1.6	0.29	0.96	0.56	1.02	25.66
b=1.5	0.35	0.90	0.42	1.15	31.33

ur: unemployment rate; M: the mass of active firms; E: total income
$\sigma = 3.8$, $\gamma = 0.6$, $\psi = 0.3$

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

	Welfare	Ave(wage)	Ave(ur)	M	E
b=1.6	0.38	0.95	0.41	2.61	53.49
b=1.5	0.29	0.90	0.51	1.90	41.07

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

5 Trade Liberalization and Firm Organization

In this section, I extend the baseline model into the international context by considering two symmetric countries. My analysis of opening up to trade in the symmetric two-country case follows Melitz (2003). I make the standard assumption that there is 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 3. The variable trade cost implies 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 such that there is selection into exporting in the CES sector. The homogeneous good is not traded regardless of the trade costs, because the two countries are symmetric.

This section focuses on how firms respond to trade liberalization. The analysis is motivated in part by recent empirical evidence on how the internal organization of firms evolves after bilateral trade liberalization. Guadalupe and Wulf (2010) showed that after the enactment of the NAFTA, American firms in sectors with larger reductions in import tariffs 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 received more incentive-based pay after the CEO increased the span of control. The extended model presented in this section rationalizes these findings.

In a world 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

$$q_d = \frac{q\left(\frac{A_H}{A_F}\tau^{\beta}\right)^{\sigma}}{1+\left(\frac{A_H}{A_F}\tau^{\beta}\right)^{\sigma}}, \tag{33}$$

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

$$\left(1+\frac{1}{\tau^{\sigma-1}}\right)^{\frac{1}{\sigma}}A, \tag{34}$$

where $A \equiv A_H = A_F$.

The equilibrium conditions in the open economy are similar to those derived in the closed economy. First, the product-market-equilibrium conditions now involve four equations. The equilibrium condition that pins down the cutoff for exporting (i.e., $\bar{\theta}_x$) is

$$\Pi(\bar{\theta}_x, (1+\frac{1}{\tau^{\sigma-1}})^{\frac{1}{\sigma}}A) - \Pi(\bar{\theta}_x, A) = f_x. \tag{35}$$

Note that the cutoff for exporting *cannot* be solved analytically, since the average cost is endogenously determined and depends on the number of layers the firm has. Next, the FE condition now becomes

$$\int_{\bar{\theta}}^{\bar{\theta}_x} \Pi(\theta, A)g(\theta)d\theta + \int_{\bar{\theta}_x}^{\infty} \Pi(\theta, (1+\frac{1}{\tau^{\sigma-1}})^{\frac{1}{\sigma}}A)g(\theta)d\theta = f_e, \tag{36}$$

since non-exporters and exporters face different adjusted market sizes. And, the ZCP condition is the same as in the closed economy, or

$$\Pi(\bar{\theta}, A) = 0. \tag{37}$$

Finally, the market-clearing condition for the homogeneous sector is still given by

$$p_h L_h = (1-\gamma)E. \tag{38}$$

Second, the equilibrium conditions for the labor markets are similar to those derived in the closed economy except that labor demand now contains two parts now: one from non-exporting firms and the other one from exporting firms. More specifically, the number of workers who choose to enter the CES sector (i.e., L_c) can be derived from the workers' indifference condition, or

$$\begin{aligned} L_c &= \int_{\theta=\bar{\theta}}^{\infty} \sum_{i=1}^{T(\theta,A)} Q(\theta, i)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta \\ &= \frac{WP(\bar{\theta}, A, M) - \psi LD(\bar{\theta}, A, M)}{p_h}, \end{aligned} \tag{39}$$

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, or

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

Note that L_c and LD now consist of two parts and are affected by the trade costs (i.e., τ and f_x). Next, the labor-market-clearing condition indicates that the number of workers

employed in the homogeneous sector is the difference between the labor endowment and the number of workers who choose to enter the CES sector, or

$$L_h = L - L_c. \tag{40}$$

Finally, the market-clearing condition of the final composite good is modified to

$$\begin{aligned} E &= \int_{\bar{\theta}}^{\bar{\theta}_x} LC(\theta, A)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta + \int_{\bar{\theta}_x}^{\infty} LC(\theta, (1+\frac{1}{\tau^{\sigma-1}})^{\frac{1}{\sigma}}A)\frac{Mg(\theta)}{1-G(\bar{\theta})}d\theta \\ &\quad \Big[f_0 + f_x\big(\frac{\bar{\theta}}{\bar{\theta}_x}\big)^k + f_e\big(\frac{\bar{\theta}}{\theta_{min}}\big)^k\Big]M + \psi M, \end{aligned} \tag{41}$$

since exporting firms use the final composite good to pay the fixed trade cost and face a different market size than exporters. The equilibrium in the open economy is characterized by seven equations (i.e., equations (35) to (41)) and seven endogenous variables (i.e., $\bar{\theta}$, $\bar{\theta}_x$, M, p_h, L_c, L_h and E). It is easy to prove that there exists a unique equilibrium under restrictions on parameter values using the same approach as the one used in the closed economy. I omit the proof to save space.

I analyze how the opening up of trade affects the internal organization of firms. The key to understanding why the the opening up of trade brings about a 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 shrinks, while the adjusted market size faced by exporters increases. Furthermore, the exit cutoff for the quality draw increases.*

Proof: See Appendix 9.3.10.

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-exporters flatten their hierarchies by reducing the number of layers and increasing the span of control at existing layers, while exporting firms increase the number of layers and reduce the span of control at all existing layers. The following proposition summarizes these results.

Proposition 9 *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 at existing layers when the number of layers is reduced. Exporting firms increase the number of layers weakly and reduce the span of control at existing layers when a new layer is added. Non-exporters increase the amount of incentive-based pay when they de-layer.*

Proof: See Appendix 9.3.11.

The effect of bilateral trade liberalization on the internal organization of firms is heterogeneous and depends on the situation firms face. In the theory, non-exporting firms

reduce the number of layers and increase the span of control at existing layers due to the shrinking market size after bilateral trade liberalization. This is what Guadalupe and Wulf (2010) found for American firms in industries with increasing import competition. Furthermore, according to the theory, firms increase the incentive-based pay 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 bilateral trade liberalization. Guadalupe and Wulf (2010) did 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 summary, firms facing different changes in the adjusted market size change their internal organization differently after bilateral trade liberalization.

6 Management Quality and Aggregate Trade Variables

In this section, I investigate how management quality affects the trade share and the WGT. I treat the number of layers as a continuous variable à la Keren and Levhari (1979) and Qian (1994) in order to derive analytical results in this section.[35] As I have shown before, the qualitative results of the model at the *aggregate-level* are unchanged, if the number of layers is treated as a continuous variable instead of a discrete variable. I derive three main results. First, the WGT are not guaranteed due to the existence of multiple frictions in the economy. Second, the trade share is shown to be bigger between economies with better MT. Third, I show that under certain conditions, WGT do exist, and economies with superior MT benefit disproportionately from the opening of trade. Finally, I relate my result to the ACR formula and argue that information on *micro-level* variables which are related to the management quality is needed for the evaluation of the WGT. In particular, the two aggregate trade statistics that appear in the ACR formula are *not* enough for us to calculate the WGT.

I state several analytical results here, and readers are referred to Appendix 9.5 for details. First, the relationship between the exporting cutoff and the exit cutoff is

$$\bar{\theta}_x = \bar{\theta}\frac{f_x \tau^{\sigma-1}}{(f - ge)}, \tag{42}$$

where $g \equiv b\psi$ and $e = 2.71828$ is Euler's number. Second, the domestic consumption share and the import share of the CES goods are

$$\lambda(\tau, b) = \frac{\tau^{\sigma-1}\left(\frac{\bar{\theta}_x}{\bar{\theta}}\right)^{k-1}}{1 + \tau^{\sigma-1}\left(\frac{\bar{\theta}_x}{\bar{\theta}}\right)^{k-1}} \tag{43}$$

and

$$1 - \lambda(\tau, b) = \frac{1}{1 + \tau^{\sigma-1}\left(\frac{\bar{\theta}_x}{\bar{\theta}}\right)^{k-1}} \tag{44}$$

[35] The analysis in all other sections treats the number of layers as a discrete variable.

respectively. Since $\frac{\bar{\theta}_x}{\bar{\theta}}$ increases in b from equation (42), the better is the MT, the bigger is the share of exporting firms. Furthermore, equation (44) shows that the better is the MT, the larger is the trade share.[36] In economies with better MT, the exit cutoff for the demand draw is higher when these economies are in autarky as a result of the selection effect. This leads to the result that active firms in these economies are bigger and more productive when these economies are in autarky. Since it is the bigger firms that start to export when the economies move from autarky to open economies, disproportionately more firms export in economies with better MT.

Next, I discuss sufficient conditions under which the WGT exist and economic interpretations of these conditions. Detailed proof can be found in Appendix 9.5. There are two sets of conditions that assure the WGT, and two crucial parameters that determine the existence of the WGT are the management quality parameter (i.e., $\frac{1}{b}$) and the elasticity of substitution (i.e., σ).

First, when the elasticity of substitution is big, there are WGT if the management quality is high. When the iceberg trade cost goes down, it induces resource reallocation between sectors. When the elasticity of substitution is high, resources are reallocated from the homogeneous sector to the CES sector. There are two countervailing forces that cause this result. First, average firm size in the CES sector *increases* due to a reduction in the iceberg trade cost. Second, the mass of active firms in the CES sector *decreases* when the iceberg trade cost goes down. The relative strength of these two forces depends on the elasticity of substitution (and the relative size of the CES sector). When the elasticity is high (i.e., smaller market power in the CES sector), the increase in average firm size is bigger, and the decrease in the mass of active firms is smaller. Therefore, resources are reallocated from the homogeneous sector *to the CES sector* after a reduction in the iceberg trade cost. I.e., the monopolistically competitive sector expands as a result of the trade liberalization. When the management quality of firms in the CES sector is high, resources that are reallocated into the CES sector are used efficiently after the trade shock, which ensures WGT.[37]

Second, when the elasticity of substitution is small, there are WGT if the management quality of firms in the CES sector is low. The economic reasoning follows the same logic that I have discussed above. Namely, the increase in average firm size is dominated by the decrease in the mass of active firms in the CES sector, when the iceberg trade cost falls and the elasticity of substitution is small. As a result, resources are reallocated *from the CES sector* to the homogeneous sector after a reduction in the iceberg trade cost. When the management quality of firms in the CES sector is low, resources that are reallocated away from the CES sector are used inefficiently before the trade liberalization. Therefore, there are WGT when these resources are reallocated to the homogeneous sector after the trade shock. In summary, WGT are not guaranteed in a world with multiple frictions. In particular, whether or not there are WGT crucially depends on whether or not resources that are reallocated after the trade liberalization are used inefficiently before the trade liberalization.

In Appendix 9.5, I show that under the first set of the above conditions, the complementarity result holds. Namely, the better is the MT, the larger are the WGT. I focus on

[36]Note that this result still holds if I define trade share as imports divided by the total income, since expenditure on the CES goods is a constant fraction of the total income.

[37]Remember that there is no incentive problem for firms in the homogeneous sector.

this set of conditions, since the parameters of the calibrated model presented in Section 7 satisfy these conditions. In short, the WGT and the complementarity result are shown to exist for counterfactual experiments that I am going to present in the next section.

Finally, I relate the formula for the WGT derived in this subsection to the formula derived in Arkolakis, Costinot and Rodriguez-Clare (2012) (henceforth, ACR formula). In an influential paper, Arkolakis, Costinot and Rodriguez-Clare showed that the WGT are *completely* determined by two aggregate statistics in a number of canonical trade model (e.g., Krugman (1980), Eaton and Kortum (2002) and Melitz (2003)). They are the share of expenditure on domestic goods and the elasticity of trade with respect to variable trade costs. In this subsection, I show that the model presented in this paper *does not* fall into the set of models whose predictions for the WGT are completely characterized by the above two statistics. Furthermore, I show that micro-level information which is related to the management quality is needed for the evaluation of the WGT.

Among the two aggregate statistics appearing in the ACR formula, one statistic (i.e., the domestic consumption share) is derived in equation (43). The other one which is the elasticity of the trade share with respect to the variable trade cost is calculated as

$$\epsilon \equiv \frac{\partial \ln\left(1-\lambda(\tau,b)\right)/\lambda(\tau,b)}{\partial \ln \tau} = -(\sigma-1)k.$$

Note that this elasticity is constant in the case of the continuous number of layers and does not depend on the management quality. However, it should be emphasized that this result is valid, only when I treat the number of layers as a continuous variable.

Now, I evaluate the WGT. Calculation shows that the WGT are

$$WGT(b,\tau) = \lambda(\tau,b)^{\frac{\gamma}{k(\gamma-(1-\gamma)(\sigma-1))}} \left[\frac{\left[1+\frac{(1-\gamma)\sigma}{\gamma(\sigma-1)}\frac{be}{(be-1)}\right] x_T(\bar{\theta}) - \frac{k-1}{k}\lambda(\tau,b)}{\left[1+\frac{(1-\gamma)\sigma}{\gamma(\sigma-1)}\frac{be}{(be-1)}\right] x_T(\bar{\theta}) - \frac{k-1}{k}} \right]^{\frac{\gamma}{\gamma-(1-\gamma)(\sigma-1)}}, \quad (45)$$

where $x_T(\theta)$ is the output level of the smallest firms in equilibrium. It is evident from equation (45) that information on the expenditure share on the CES goods (i.e., γ), the management quality (i.e., $\frac{1}{b}$), and the output level of the smallest firms (i.e., $x_T(\bar{\theta})$) is needed for the evaluation of the WGT. In an extreme case in which there is no outside sector, the above equation is simplified to

$$WGT(b,\tau) = \lambda(\tau,b)^{\frac{1}{k}} \frac{x_T(\bar{\theta}) - \frac{k-1}{k}\lambda(\tau,b)}{x_T(\bar{\theta}) - \frac{k-1}{k}}.$$

Even in this extreme case, information on three variables is still needed to evaluate the welfare change from opening up to trade. A new variable that does not show up in the ACR formula is the output level of the smallest firms in equilibrium.

In summary, in a world with information frictions which validate the use of management hierarchies, evaluating the WGT requires more information than that contained in the ACR formula. Of course, this does not mean that the ACR formula is incorrect, since one restriction of the ACR formula that aggregate profit is a constant fraction of total income is violated in my framework. The above result only shows that micro-level factors such as the management quality should be taken into account, even if we only care about changes in *aggregate* economic outcomes.

7 Calibration

In this section, I calibrate the model in order to implement counterfactual experiments. The counterfactual experiments show that an improvement in MT has quantitatively important effects on aggregate productivity, firm size and welfare. Furthermore, its impact on the WGT is quantitatively sizable.

I choose to match six moments from the data in order to estimate six parameters: $b\psi$, k, f_e, f, f_x, and γ. The first four moments are moments that are not related to international trade. They are the average employment of firms, the fraction of firms that have less than ten employees, the fraction of firms that have more than five hundred employees, and the Pareto shape parameter obtained from the regression of log rank on log firm size (i.e., log employment).[38] The other two moments are related to international trade. They are the share of firms that export and the trade share (i.e., the average of the export-GDP ratio and the import-GDP ratio).

I obtained values for the above moments from various sources. First, from the U.S. economic census data, I obtained values of three moments. In 2007, an average U.S. manufacturing firm employs 46.57 workers. There are 55.04% manufacturing firms that have less than ten employees, and 1.03% manufacturing firms that have more than five hundred employees in 2007. Second, the world bank database shows that the export-GDP ratio and the import-GDP ratio are 12% and 16% for the U.S. in 2007 respectively. Thus, the trade share for the U.S. is 14% in 2007. Bernard et al. (2007) estimated that 18% of U.S. manufacturing firms exported in 2002.[39] Finally, Caliendo and Rossi-Hansberg (2012) reported that the Pareto shape parameter for U.S. manufacturing firms is 1.095. For details, see Table 3.

I obtained values of the other parameters from the literature. Following Bernard et al. (2003) and Melitz and Redding (2013), I set the elasticity of substitution, σ, to 4. Following Caliendo and Rossi-Hansberg (2012), I set the labor endowment, L, to 16.48 million. Following Ghironi and Melitz (2005), I set the iceberg trade cost, τ, to 1.3.

I search over the parameter space for the parameter values that match the above moments, using as a loss function the norm of the percentage deviation difference between the model and the data.[40] Estimated parameters are reported in Table 4. The estimation of ψ uses the result from Hall (2006) that the disutility of working is 71% of average labor productivity (i.e., output per worker). Since the estimated value of $b\psi$ is 0.649, the implied ψ and b are 0.370 and 1.755 respectively. As an over-identification check, I obtained values for two additional moments that are closely related to the internal organization of firms to check whether the calibrated model does a good job at matching these two moments as well. First, the U.S. economic census data shows that wage share of production workers was 57.7% in 2007. As Table 5 shows, the calibrated model predicts that this value is 52.1% which is close to the data.[41] Second, Guadalupe and Wulf (2010)

[38]The Pareto shape for the firm size distribution is obtained from the regression of $\ln(Pr(employment > x))$ on $\ln(employment) = x$, where $Pr(employment > x)$ is the share of firms that have employment more than x.

[39]No information is available for this moment in 2007.

[40]In other words, I use the identity matrix as the weighting matrix.

[41]Although non-production workers only monitor in my model, it is straightforward to extend the model into the case in which non-production workers both monitor and exert effort to increase firm productivity under certain assumptions. If I assume that monitoring requires only time and no effort as

reported that the average number of managerial layers was $3.1-3.2$ for various industries in 1999 in a sample containing 230 large U.S. manufacturing firms.[42] Although I use data on the entire population of manufacturing firms in 2007, it is interesting to compare the characteristics of firms in their sample relative to what my model yields. The calibrated model predicts that the average number of managerial layers is 3.44 which is close to what Guadalupe and Wulf (2010) have found. In summary, the difference between the moments obtained from the data and the moments generated by the calibrated model is small.

Table 3: Moments from the Data and the Model

	Data	Model
Pareto Shape Parameter	1.095	1.097
Employment per firm	46.57	46.57
Fraction of small firms (less than 10 employees)	55.04%	55.14%
Fraction of big firms (more than 500 employees)	1.03%	0.76%
Fraction of exporting firms	18%	17.84%
Trade Share (relative to GDP)	14%	14.00%

The Pareto shape parameter is obtained from the regression of log rank on log firm size.

Table 4: Parameter Values

	Value	Sources
σ	4	Bernard et al. (2003)
L	16.48	Caliendo and Rossi-Hansberg (2012)
τ	1.3	Ghironi and Melitz (2005)
f	4.98	
f_e	3.00	
f_x	4.84	
k	1.188	
$b\psi$	0.649	
γ	0.545	

Table 5: Moments Related to Internal Firm Organization

	Data	Model
The number of managerial layers	3.1-3.2	3.44
Wage share of production workers	57.7%	52.1%

in Qian (1994), the extend model yields exactly the same firm-level outcomes and resource allocation as the current model. Therefore, the value of this moment in the data is comparable to the value of this moment generated by the model.

[42]The number of managerial layers is the number of layers ranging from the CEO to division managers who supervise the lowest production units. In the model, it is measured as the number of layers ranging from layer $T-1$ to layer 0.

I implement two counterfactual experiments to investigate the quantitative impact of an improvement in MT on firm size, aggregate productivity and the WGT. The average score of Indian firms on monitoring and incentive practices is about 22% lower than American firms in the World Management Survey (WMS). I consider a similar percentage change in the management quality in the model. More specifically, I consider a comparative statistics in which the value of b decreases from $2.141(= 1.755 * 1.22)$ (India) to 1.755 (the US). Remember that the management quality is denoted by $\frac{1}{b}$. Thus, such an improvement implies a 22% improvement in MT. The first quantitative exercise is to evaluate how such an improvement affects firm size, productivity and welfare. Table 6 shows that average employment and sales increase by 72.2% and 36.9% as a result of this improvement. Furthermore, weighted average firm productivity and welfare increase by 22.2% and 47.1% respectively. The second quantitative exercise is to evaluate how the deterioration in MT affects the trade share and the WGT. Table 7 reports the result. The improvement in MT yields a 1.21% increase in the WGT which is about 7.4% of the WGT before the improvement. This improvement also generates a 5.31% increase in the fraction of exporting firms. In total, the calibrated model is able to generate a quantitatively sizable impact of an improvement in MT on firm size, productivity and the WGT.

Table 6: Firm Size, Aggregate Productivity and Welfare

	$Prod_w$	$Employment$	$Sales$	$Welfare$
$M\&I \uparrow 22\%$	22.2%	72.2%	36.9%	47.1%

$Prod_w$: percentage change in weighed average of firm productivity
$Employment$: percentage change in average employment
$Sales$: percentage change in average sales
$Welfare$: percentage change in welfare

Table 7: Trade Share and the Welfare Gains from Trade

	b=2.141	b=1.755	Complementarity
WGT	16.42%	17.63%	1.21%
Share of exporting firms	12.53%	17.84%	5.31%

8 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 findings related to changes in the organizational structure of firms after

bilateral trade liberalization, but can also discuss how the quality of MT affects aggregate trade variables such as the trade share and the WGT.

The main contribution of this paper is to explore the selection effect of a common improvement in MT for firms. This effect is due to the heterogeneous impact of such an improvement on firms with various efficiency levels. As a result of this selection effect, resources are reallocated from small firms to big firms, which leads to systematic changes in firm size, aggregate productivity, and welfare. It is for the first time in the literature that these economic insights and implications are pointed out rigourously in a general equilibrium framework. Furthermore, the channel through which the common improvement in MT generates the selection effect is shown to be the endogenous formation of the internal firm organization. These economic insights and implications open room for future research on management practices and should be contrasted with the data.

Undoubtedly, much more research remains to be done. First, integrating the knowledge-based hierarchy and the incentive-based hierarchy into a unified framework is an interesting idea, since each approach reflects only one part of the function of the management hierarchy. Second, investigating how the quality of MT affects firms' organizational choice (e.g., outsourcing or in-house production) is also an interesting topic to explore. Finally, although MT considered in this chapter is exogenous and invariant across firms, recent evidence suggests that MT does differ across firms and is affected by shocks such as trade liberalization (Bloom and Van Reenen (2010) and Bloom, Sadun, and Van Reenen (2012b)). Therefore, it is worth exploring how trade liberalization *endogenously* affects MT which in turn impacts aggregate economic outcomes.

9 Appendix

9.1 Empirical Motivation

In this subsection, I discuss the details of my empirical motivation. Three parts come in order. First, I discuss the content of MT used in this paper. Next, I argue that MT is highly correlated with firm performance by showing some motivating evidence. 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 with 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 the concept of MT used in this paper. Among the seven practices I pick up, the first four items are "performance tracking", "performance review", "performance dialogue" and "performance clarity", which are related to whether or not firms can successfully find and catch misbehaving employees. The other three items are "consequence management", "rewarding high performance", "remove poor performers", which 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 is highly correlated with firm performance by presenting simple scatter plots in Figure 7. As it is evident in the figure, the quality of MT is positively associated with firm performance such as sales per employee or total employment.

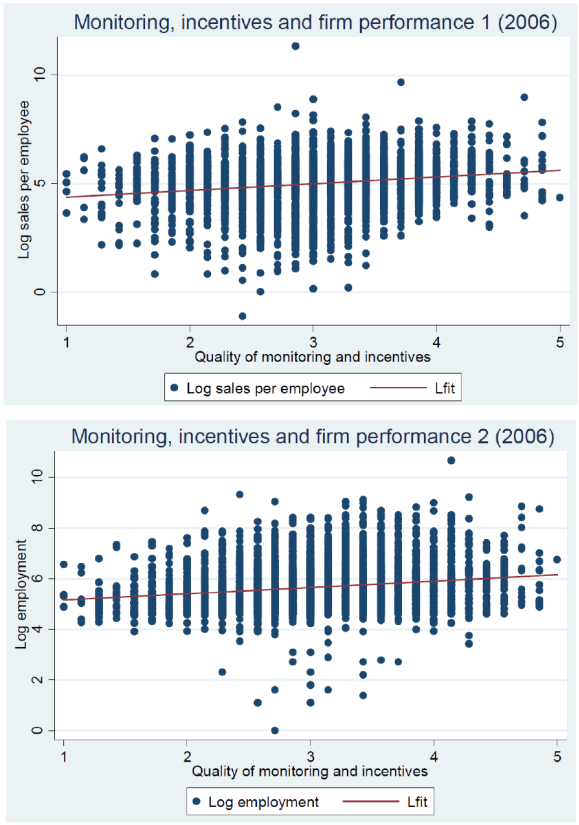
Third, I implement a variance and covariance decomposition exercise and show that there is substantial heterogeneity for scores on monitoring and incentives across firms. Note that the average management score is the sum of two parts or

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

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. Next, I decompose the variation in the average management score into the following three parts:

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

Figure 7: 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 8 shows that there is substantial variation in the score of *moinc* across firms.

Table 8: Management Score Decomposition

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

moinc: average score on monitoring and incentives.

nonmoinc: average score on other management practices.

crossterm: correlation between the above two scores across firms.

Finally, I show that quality of monitoring and incentives differs across countries and is associated with firm size distribution. First, Table 9 shows that scores on every management practice that is included in the set of *moinc* differ significantly between China and the U.S. Second, Figure 8 shows that the distribution of scores on *moinc* differs substantially between India and the U.S.[43] Average score on *moinc* is significantly higher for U.S. firms. Finally, average management score on *moinc* is positively associated with the average firm size in an economy as Figure 8 shows. In other words, U.S. firms that have good MT are larger than Indian firms on average, and there are much more small firms in India compared with the U.S.

Table 9: 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: $1 - 5$.

All differences in means are statistically significant at 1% level.

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

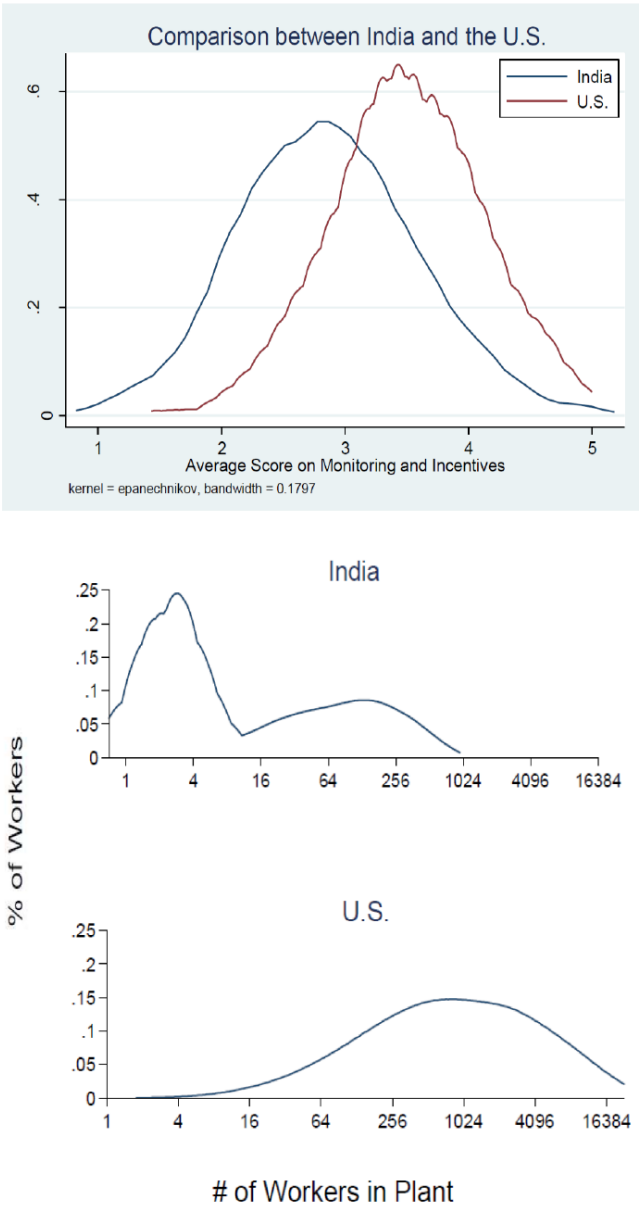
9.2 Continuous Effort Choice

In this subsection, I setup up the model under the assumption that workers' effort choice is a continuous variable. The goal of this exercise is to show that the model under this alternative assumption is isomorphic to the model with a binary effort choice.

I discuss the wage determination under the alternative assumption for the effort choice. Now, non-production workers' effort choice a_i ($i \leq T-1$) is a continuous variable with an upper bound $\bar{a}$. An increase in a_i raises the probability of catching and firing a

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

Figure 8: Management Quality and the Firm Size Distribution



misbehaving worker even when the span of control is unchanged. Under this specification, the worker's incentive compatibility constraint is

$$w_i - \psi(a_i) \geq (1 - p(b, x_{i-1}, a_{i-1}))w_i \tag{46}$$

where $p(b, x_i, a_{i-1}) = \frac{1}{b}\frac{a_{i-1}}{x_{i-1}} = \frac{1}{b}\frac{a_{i-1}m_{i-1}}{m_i}$. Thus, the incentive compatible wage is

$$w_i = \frac{b\psi(a_i)m_i}{a_{i-1}m_{i-1}}. \tag{47}$$

For reasons of tractability, I choose the following functional form for the cost function of exerting effort:

$$\psi(a) = \frac{\bar{\psi}}{\bar{a} - a}. \tag{48}$$

The key feature of the above functional form is that disutility increases with the effort level at an increasing speed (i.e., a convex function).

I show why a firm wants to incentivize workers at a given layer to choose the same level of effort and allocates the monitoring intensity evenly across them. The basic logic is the same as in the discrete case. The firm can always reduce the wage payment *and* induce the same total amount of effort inputs by equalizing the effort level and the monitoring intensity across workers at a given layer. Lemma 3 summarizes the results.

Lemma 3 *The firm incentivizes workers at a given layer to choose the same level of effort and equalizes the monitoring intensity across them.*

Proof. I prove this lemma in three steps. For any two units of effort inputs at layer i and the corresponding monitoring intensities for them, (i.e., a_{i1}, p_{i1}) and (a_{i2}, p_{i2}), there are four possibilities of relationship between them in total.[44] Namely, $a_{i1} = a_{i2}$ and $p_{i1} = p_{i2}$; $a_{i1} = a_{i2}$ and $p_{i1} \neq p_{i2}$; $a_{i1} \neq a_{i2}$ and $p_{i1} = p_{i2}$; $a_{i1} \neq a_{i2}$ and $p_{i1} \neq p_{i2}$. I prove that the last three possibilities are not optimal in what follows.

First, for the case in which $a_{i1} = a_{i2}$ and $p_{i1} \neq p_{i2}$, I can use exactly the same approach used in the discrete case to prove that it is not optimal. Second, I discuss the case in which $a_{i1} \neq a_{i2}$ and $p_{i1} = p_{i2}$. Total wage payment that is used to induce effort levels a_{i1} and a_{i2} can be reduced, if the effort levels are equalized. Formally, it must be true that

$$b\frac{\psi(\frac{a_{i1}+a_{i2}}{2})}{p_{i1}} + b\frac{\psi(\frac{a_{i1}+a_{i2}}{2})}{p_{i2}} < b\frac{\psi(a_{i1})}{p_{i1}} + b\frac{\psi(a_{i2})}{p_{i2}},$$

as $\psi(a)$ is a convex function. Finally, for the case in which $a_{i1} \neq a_{i2}$ and $p_{i1} \neq p_{i2}$, let me make the following simplifying notations:

$$a_0 \equiv a_{i1} + a_{i2}$$

and

$$p_0 \equiv p_{i1} + p_{i2}.$$

[44] The monitoring intensity is defined as $p_i \equiv \frac{a_{i-1}m_{i-1}}{m_i}$.

Cost minimization requires that allocation of two units of effort inputs and the monitoring intensities given (a_0, p_0) are optimal at (a_{i1}, p_{i1}) and (a_{i2}, p_{i2}). Thus, we must have

$$\frac{\psi^{'}(a_{i1})}{p_{i1}} = \frac{\psi^{'}(a_{i2})}{p_{i2}}$$

and

$$\frac{\psi(a_{i1})}{p_{i1}^2} = \frac{\psi(a_{i2})}{p_{i2}^2}$$

due to the FOCs with respect to a_{i1} and p_{i1}. The above two equations lead to the result that

$$a_{i1} = a_{i2} = \frac{a_0}{2}$$

and

$$p_{i1} = p_{i2} = \frac{p_0}{2}$$

which contradicts that $a_{i1} \neq a_{i2}$ and $p_{i1} \neq p_{i2}$. Therefore, the cost structure that is consistent with the firm's cost minimization is the case in which $a_{i1} = a_{i2}$ and $p_{i1} = p_{i2}$ for any two units of effort inputs. In other words, the firm incentivizes workers at a given layer to choose the same level of efforts and equalizes the monitoring intensities on them. QED.

I can characterize the firm's optimization problem now. Based on equation (47) and Lemma 3, the firm's optimization problem can be written as

$$\begin{aligned} \max_{\{a_i, m_i\}_{i=1,\ldots,T}, N, T} \quad & A(\theta)^{\frac{1}{\sigma}} q^{\frac{\sigma-1}{\sigma}} - \sum_{i=1}^{T} b\psi(a_i) \frac{m_i^2}{a_{i-1} m_{i-1}} \\ s.t. \quad & x_i = \frac{m_i}{m_{i-1}}, \\ & x_0 = 1;\ a_0 = 1;\ m_T = N = \frac{q}{a_T};\ m_0 = 1; \\ & 0 \leq a_i \leq \bar{a}. \end{aligned} \tag{49}$$

As in the discrete case, I minimize the firm's variable cost given an output level first and then solve for the optimal output level as well as the optimal number of layers.

First, the variable cost for a firm that produces q units of output and has $T+1$ layers is

$$\min_{\{a_i, m_i\}_{i=1,\ldots,T}} \sum_{i=1}^{T} b\psi(a_i) \frac{m_i^2}{a_{i-1} m_{i-1}}, \tag{50}$$

where $m_T = q/a_T$. The FOC of equation (50) with respect to m_i is

$$\frac{2m_i \psi(a_i)}{a_{i-1} m_{i-1}} - \frac{\psi(a_{i+1}) m_{i+1}^2}{a_i m_i^2} = 0. \tag{51}$$

The FOC of equation (50) with respect to a_i is

$$\frac{m_i^2 \psi^{'}(a_i)}{a_{i-1} m_{i-1}} - \frac{\psi(a_{i+1}) m_{i+1}^2}{a_i^2 m_i} = 0. \tag{52}$$

Note that the second order conditions hold for both a_i and m_i. By comparing equation (51) with equation (52), I have

$$\frac{2\psi(a_i)}{m_i\psi'(a_i)} = \frac{a_i}{m_i}$$

or

$$\frac{\psi'(a_i)a_i}{\psi(a_i)} = 2.$$

Due to the functional form of $\psi(a)$ given in equation (48), the optimal effort level a_i $(i < T)$ is

$$a_i^* = \frac{2\bar{a}}{3}. \tag{53}$$

Without loss of generality, I normalize $\bar{a}$ to $\frac{3}{2}$ which leads to

$$a_i^* = 1$$

for all $i < T$. Thus, the total wage payment given q and T is

$$\min_{a_T, \{m_i\}_{i=1,\ldots,T}} \sum_{i=1}^{T-1} b\psi(1)\frac{m_i^2}{m_{i-1}} + b\psi(a_T)\frac{(\frac{q}{a_T})^2}{m_{T-1}}, \tag{54}$$

from which I solve for optimal a_T. Using the FOC of equation (54) with respect to a_T, I derive that

$$\frac{\psi'(a_T)a_T}{\psi(a_T)} = 2 \tag{55}$$

and

$$a_T^* = \frac{2\bar{a}}{3} = 1.$$

In total, the optimal effort choice is one for *all* workers, which is the same as in the discrete case.[45]

Second, the FOCs of equation (50) with respect to various m_i's can be rewritten as

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

which are the same as the ones derived in the discrete case. Therefore, I obtain the same solution for optimal employment at layer i as in the discrete case or

$$m_i(N,T) = 2^i\Big(\frac{N}{2^T}\Big)^{\frac{2^T-2^{T-i}}{2^T-1}}. \tag{56}$$

Third, I derive optimal output and operating profit for firms with the quality draw θ from equation (49). The solutions are

$$q(\theta,T) = N(\theta,T) = \left[\frac{A\beta\theta^{\frac{1}{\sigma}}}{b\psi 2^{2-\frac{T}{2^T-1}}}\right]^{\frac{\sigma(2^T-1)}{\sigma+(2^T-1)}} \tag{57}$$

[45]The key equations here are equations (52) and (55). The key feature of them is that the optimal effort choice is *completely* determined by the cost function to exert effort. Thus, neither firm-level characteristics nor market-level variables are needed when we derive it.

and

$$\pi(\theta,T)=(1-\frac{\beta(2^T-1)}{2^T})(A\theta^{\frac{1}{\sigma}})^{\frac{2^T\sigma}{\sigma+(2^T-1)}}\left(\frac{\beta/b}{\left(2^{\frac{2^{T+1}-2-T}{2^T-1}}\right)}\right)^{\frac{(\sigma-1)(2^T-1)}{\sigma+(2^T-1)}}, \tag{58}$$

These two solutions are exactly the same as the ones derived in the discrete case.

Finally, the firm chooses the number of layers optimally. The proofs and results in the discrete case all apply here, since the cost structure derived in the continuous case is exactly the same as the one derived in the discrete case. Furthermore, all empirical predictions and propositions derived in the discrete case hold in the continuous case as well.

9.3 Proof

9.3.1 Proof of Lemma 1

Proof. First, suppose that 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. 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 the labor costs, 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 intensities evenly across workers is that it could reduce wage payments by doing so, if the monitoring intensities were not equalized. More specifically, suppose there are two units of effort inputs 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 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 intensities across these two units of effort 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 elicit the two units of effort inputs under a lower cost. Therefore, the firm's optimal choice is to equalize the monitoring intensities across workers at a given layers. QED.

9.3.2 Proof of Proposition 1

The AVC function and the MC function given the number of layers are

$$AVC(q,T)=(2-\frac{1}{2^{T-1}})b\psi 2^{1-\frac{T}{2^T-1}}q^{\frac{1}{2^T-1}} \tag{59}$$

and

$$MC(q,T) = b\psi 2^{2-\frac{T}{2^T-1}} q^{\frac{1}{2^T-1}} = \frac{2^T}{2^T-1} AVC(q,T). \tag{60}$$

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 4 *q_T increases in T.*

Proof. I 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 4, I can characterize the overall shape of the AVC curve.

Lemma 5 *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 5. 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 5 and Claim 1. QED.

9.3.3 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*

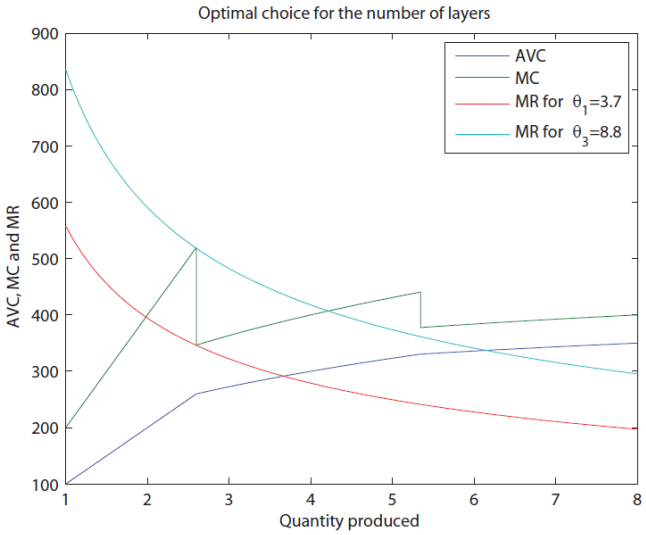
$$MR(\theta_{T1}, q_T) = A\beta\theta_{T1}^{\frac{1}{\sigma}} q_T^{-\frac{1}{\sigma}} = MC(q_T, T+1) = b\psi 2^{2-\frac{T+1}{2^{T+1}-1}} q_T^{\frac{1}{2^{T+1}-1}}.$$

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.

Figure 9: Lower and Upper Bounds on Layer-Switching from $T = 1$ to $T = 2$



The graphical representation of θ_{T1} and θ_{T3} is in Figure 9.

Second, I show that only when the firm's quality draw is between $[\theta_{T1}, \theta_{T3}]$, does it have incentive to switch from $T+1$ layers to $T+2$ layers in the following lemma.

Lemma 6 *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.*

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

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

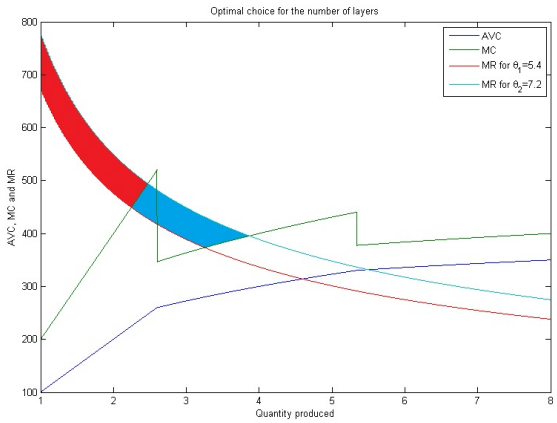
where I have used the result that $AVC(q_T, T) = AVC(q_T, T+1)$. Similarly, when $\theta = \theta_{T3}$, choosing $T+2$ layers yields more profit as

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

QED.

Third, I use the following lemma to show the complementarity between the benefit of adding a layer and the quality draw θ.

Figure 10: Optimal Choice of the Number of Layers



Lemma 7 *For a given T, $\pi(\theta, T+1) - \pi(\theta, T)$ increases continuously in θ for $\theta \in [\theta_{T1}, \theta_{T3}]$.*

Proof. I use Figure 10 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

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

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.

Fourth, 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 3 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 6, 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 7, 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 3 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 3, 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 become much less time-consuming because of this result.

9.3.4 Proof of Proposition 3

First, when the firm expands without changing the number of layers, both employment and output increase continuously due to equations (13) and (15). Second, the firm's optimal pricing rule implies that

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

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

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

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.

9.3.5 Proof of Proposition 4

First, equation (14) 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 (9) is positively affected the span of control, wages increase at all layers. Third, the FOCs with respect to employment in equation (12) 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.[46] 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:[47]

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

This is because

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

[46] This result will be used later.

[47] q_T is defined in Definition 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. QED.

9.3.6 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 (13) and (15), 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, what 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

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

As $\frac{q_{T0}}{2^{T0}} > 1$ for $T0 \geq 1$ due to the proof in Appendix 9.3.5, 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 the 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 the 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.

9.3.7 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

$$\frac{\partial UC_T(q,b)}{\partial q} = \frac{1}{2^{T-1}} b\psi 2^{1-\frac{T}{2^T-1}} q^{\frac{-(2^T-2)}{2^T-1}} - \frac{f_0}{q^2}. \tag{61}$$

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:

$$q_{Tm}(b) = \Big[\frac{f_0}{4b\psi}\Big]^{\frac{2^T}{2^T-1}} 2^T. \tag{62}$$

A 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

$$AVC_{T+1}(q_{T+1,m}(b),b) + \frac{f_0}{q_{T+1,m}(b)} < AVC_T(q_{T,m}(b),b) + \frac{f_0}{q_{T,m}(b)} \quad \forall T \geq 1 \tag{63}$$

One key observation is that equation (61) implies

$$\frac{1}{2^T-1}\frac{AVC_T(q_{T,m}(b),b)}{q_{T,m}(b)} = \frac{f_0}{q_{T,m}(b)^2}. \tag{64}$$

From equation (64), I conclude that

$$AVC_T(q_{T,m}(b),b) + \frac{f_0}{q_{T,m}(b)} = (2^T-1)\frac{f_0}{q_{T,m}(b)} + \frac{f_0}{q_{T,m}(b)} = 2^T\frac{f_0}{q_{T,m}(b)}. \tag{65}$$

Substituting equation (65) into equation (63) 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 (62) 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.

9.3.8 Proof of Proposition 7

The strategy to prove this proposition is the following. First, I assume that the incentive compatible wage defined in equation (9) satisfies the constraint indicated in equation (31) 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 (31) in *every* labor submarket.

First, I redefine the equilibrium using three conditions. Substituting equation (11) into equation (25) leads to the homogeneous sector's employment expressed as

$$L_h = \frac{(1-\gamma)A^\sigma P^{1-\sigma}}{\gamma p_h}. \tag{66}$$

Substituting the above equation and equation (29) into equation (29) yields the following labor market clearing condition:

$$\frac{WP(\bar{\theta}, A, M) - \psi LD(\bar{\theta}, A, M)}{p_h} + \frac{(1-\gamma)A^\sigma P^{1-\sigma}}{\gamma p_h} = L. \tag{67}$$

Now, the equilibrium of the economy can be solved using three equations (i.e., equations (23), (24) and (67)). 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, the ideal price index is

$$P = \frac{1}{p_h^{\frac{1-\gamma}{\gamma}}} \tag{68}$$

due to equation (4). Second, the ideal price index defined in equation (5) can be reexpressed as

$$P = \Big(\int_{\theta=\bar{\theta}}^{\infty} \theta p(\theta)^{1-\sigma} M \frac{g(\theta)}{1-G(\bar{\theta})} d\theta \Big)^{\frac{1}{1-\sigma}} \equiv P_1(\bar{\theta}, A) M^{\frac{1}{1-\sigma}}. \tag{69}$$

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 (68) and (69) and values of $\bar{\theta}$, A and p_h. Third, the aggregate income E can be derived by using equation (11) and value of A and P. Finally, the allocation of labor can be obtained by using equations (29) and (66) 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 (10), 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, wage payment per active firm 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.

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 (23). 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 (16) strictly increases with both θ and A. Therefore, equation (23) leads a negative relationship between $\bar{\theta}$ and A. Of course, when $\bar{\theta}$ approaches zero, A determined from equation (23) approaches infinity. And when $\bar{\theta}$ goes to infinity, A determined from equation (23) 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 (24). 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 (67) yields

$$\frac{WCper(A, \bar{\theta}) - \psi LDper(A, \bar{\theta})}{p_h} M + \frac{(1-\gamma)A^{\sigma} P^{1-\sigma}}{\gamma p_h} = L. \tag{70}$$

Next, substituting equation (69) into equation (4) leads to the expression of M in terms of p_h and $P_1(\bar{\theta}, A)$ as follows:

$$M = p_h^{\frac{(1-\gamma)(\sigma-1)}{\gamma}} P_1(\bar{\theta}, A)^{\sigma-1}. \tag{71}$$

Finally, substituting equations (69) and (71) into equation (67) results in the following labor market clearing condition:

$$\Big[WCper(A, \bar{\theta}) - \psi LDper(A, \bar{\theta})\Big] P_1(\bar{\theta}, A)^{\sigma-1} + \frac{(1-\gamma)A^{\sigma}}{\gamma} = p_h^{1-\frac{(1-\gamma)(\sigma-1)}{\gamma}} L. \tag{72}$$

There exists a unique p_h that satisfies the above equation, as long as $\frac{(1-\gamma)(\sigma-1)}{\gamma} \neq 1$.[48] Moreover, equilibrium p_h must satisfy the condition that

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

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

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 the worker who chooses to enter the CES sector earn higher expected payoff.[49] These two negative effects on the labor demand are reflected by p_h that appears in the left hand side of equation (70). 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 (9) satisfies the constraint indicated in equation (31) 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 $(\bar{\theta}, A, b)$.[50] Second, equation (72) indicates 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.

9.3.9 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.

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

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

I make the following notations. Suppose b decreases from b_1 to $b_2(< b_1)$ due to an improvement in MT. Let $\bar{\theta}_1$ (or $\bar{\theta}_2$) be the demand threshold for exiting when $b = b_1$ (or $b = b_2$). Let A_1 (or A_2) be the adjusted market size when $b = b_1$ (or $b = b_2$).

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

Lemma 8 *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 (24). 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 (10) must decrease for all firms. Thus, the exit cutoff must increase. 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 8, all surviving firms weakly increase their number of layers. 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[51]

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

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.

Fourth, 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

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

[51] 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 $A \equiv \frac{A_2}{A_1} < 1$, and $b \equiv \frac{b_1}{b_2} > 1$. This expression can be simplified further to

$$\frac{\theta_{T,2}|_{b=b_1}}{\theta_{T,2}|_{b=b_2}} = (bA)^{\sigma}. \tag{73}$$

Second, from the expression of firm's profit function derived in Equation (16), I have

$$\frac{\bar{\theta}_1}{\bar{\theta}_2} = A^{\sigma} b^{(\sigma-1)(1-\frac{1}{2^{T_0}})}. \tag{74}$$

Finally, from equations (73) and (74), 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 (74), I conclude that

$$\frac{S(t\bar{\theta}_2, A_2)}{S(t\bar{\theta}_1, A_1)} = \left[\left(\frac{1}{b}\right)^{(\sigma-1)(1-\frac{1}{2^{T_0}})} b^{\sigma}\right]^{(\sigma-1)\frac{2^{T_1}-2^{T_0}}{(\sigma+2^{T_1}-1)(\sigma+2^{T_0}-1)}} > 1. \tag{75}$$

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

$$SC_i(T, q(\theta, b, A, T(\theta, b, A))) = \frac{m_{i+1}(T, q(\theta, b, A, T(\theta, b, A)))}{m_i(T, q(\theta, b, A, T(\theta, b, A)))}, \tag{76}$$

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[52]

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

The span of control calculated in equation (14) 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.

9.3.10 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 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

9.3.11 Proof of Proposition 9

The firm's optimization problem defined 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 $\theta^{\frac{1}{\sigma}}$. It is straightforward to observe 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

[52] For detailed proof of this result, see Appendix 9.3.9.

exporters' revenue, employment, and output go up after the economy opens up to trade. As it has been shown 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. Since 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.

9.4 A Continuous Number of Layers: the Closed Economy

In this subsection, I treat the number of layers as a continuous variable à la Keren and Levhari (1979) and Qian (1994). First, I show that the main results established in the paper (i.e., Proposition 8) hold in *all* possible cases when the number of layers is treated as a continuous variable. Second, I derive a condition under which the type of equilibrium I investigate exists, and the condition only consists of exogenous parameters. Finally, I obtain qualitative results on the welfare effect of an improvement in MT.

The only change in the specifications of the model from the paper is the treatment of the number of layers. First, the effort choice is still treated as a binary variable. As a result, every worker is incentivized to exert effort in equilibrium. Second, the economy still consists of two sectors: a traditional sector and a CES sector. Third, workers can still choose which type of jobs to apply for. Based on these specifications, I can express the firm's optimization problem as

$$\begin{aligned} \max_{s_i,T} \quad & A\theta^{\frac{1}{\sigma}} x_T^{\beta} - \int_{t=0}^{T} g s_t x_t dt \\ s.t. \quad & \dot{x}_t = x_t \log(s_t) \\ & x_0 = 1, \end{aligned} \tag{77}$$

where $g \equiv b\psi$. Now, the firm chooses the span of control at each layer as well as the number of layers in equilibrium.[53] Different from the optimization problem in Qian (1994), the problem now becomes an "open-final-time" (i.e., T) and "open-end-point" (i.e., x_T) optimal control problem with one state variable x_t and one control variable s_t.

I solve this optimization problem in the following several steps. First, the corresponding Hamiltonian is

$$H(t) = -g s_t x_t + p_t x_t \log(s_t).$$

The corresponding optimality conditions are

$$\dot{p}_t = g s_t - p_t \log(s_t), \tag{78}$$

$$-g x_t + p_t x_t / s_t = 0 \tag{79}$$

and

$$H(t) = -g s_t x_t + p_t x_t \log(s_t) = 0. \tag{80}$$

[53] For more details, see Qian (1994).

Above three conditions are exactly the same as in Qian (1994), as these three conditions are sufficient and necessary conditions for an "open-final-time" and "fixed-end-point" optimal control problem. Moreover, there is an additional transversality condition for this "open-final-time" and "open-end-point" problem as follows:[54]

$$p_T = \frac{\partial A\theta^{\frac{1}{\sigma}} x_T^{\beta}}{\partial x_T} = A\beta\Big(\frac{\theta}{x_T}\Big)^{\frac{1}{\sigma}} = gs_T. \tag{81}$$

Next, based on equations (78) to (81), I derive closed-form solutions for this optimal control proble as follows:

$$s_t = e,\ w_t = ge,\ x_t = e^t,\ x_T(\theta, A) = \theta\Big(\frac{A\beta}{ge}\Big)^{\sigma},\ T = \log\theta + \sigma\log\Big(\frac{A\beta}{ge}\Big), \tag{82}$$

where e is Euler's number and equals 2.71828... Total employment is

$$l(\theta, A) = \int_{t=0}^{T} x_t dt = (x_T - 1). \tag{83}$$

The resulting operating profit and revenue are

$$\pi(\theta, A) = (1-\beta)A^{\sigma}\theta\Big(\frac{\beta}{ge}\Big)^{\sigma-1} + ge \tag{84}$$

and

$$S(\theta, A) = A^{\sigma}\theta\Big(\frac{\beta}{ge}\Big)^{\sigma-1} \tag{85}$$

respectively. The operating profit function in equation (84) is similar to the standard operating profit function in models with the CES preference except that it includes an additional term (i.e., ge). This implies that the Free Entry (FE) condition of the current model is the same as the in Melitz (2003) except that the *implied* fixed cost now becomes $f - ge$.[55]

Third, I prove the main result of this subsection. As in the paper, I still assume that the quality draw follows a Pareto distribution with parameter k.

$$G(\theta) = 1 - \Big(\frac{\theta_{min}}{\theta}\Big)^{k},$$

where $G(\theta)$ is the Cumulative Distribution Function (CDF) of θ.[56]

Proposition 10 *When the quality of MT improves in an economy (i.e., b decreases), the exit cutoff for the quality draw increases. On top of that, the firm size distribution (i.e., sales or output or employment) and the distribution of the number of layers move to the right in the FOSD sense.*

[54]See Meagher (2003) for more details.

[55]Throughout the this section, I assume that $f > ge$.

[56]The Pareto assumption is needed only for the proof of the change in distributions. The result on the change of the exit cutoff is independent of this distributional assumption.

Proof: First, the key observation is that the ratio of operating profit of a firm with a higher quality draw to operating profit of a firm with a lower quality draw increases when the quality of MT improves. More specifically, for $\theta_2 > \theta_1$ I have

$$\frac{\pi(\theta_2, A)}{\pi(\theta_1, A)} = \frac{(1-\beta)A^{\sigma}\theta_2\left(\frac{\beta}{ge}\right)^{\sigma-1} + ge}{(1-\beta)A^{\sigma}\theta_1\left(\frac{\beta}{ge}\right)^{\sigma-1} + ge}$$

decreases in g. As $g = b\psi$ decreases when the quality of MT improves, this ratio must increase when the quality of MT increases. Therefore, the exit cutoff must increase when b decreases.

Next, I discuss what happens to firms at the exit cutoff. I make the the following simplifying notations before the discussion. Let $\bar{\theta}_1$ and $\bar{\theta}_2 (> \bar{\theta}_1)$ be the exit cutoffs before and after MT improves respectively. Let A_1 and A_2 be the corresponding adjusted market size. Let b_1 and $b_2 (< b_1)$ be the inefficiency of MT before and after MT improves respectively. I prove that firms on the exit cutoff increase their revenue, employment, output and the number of layers after MT improves. First, the revenue must increase for firms at the exit cutoff as

$$S(\bar{\theta}_1, A_1) = \sigma(f - g_1 e) < \sigma(f - g_2 e) = S(\bar{\theta}_2, A_2).$$

Second, output that equals x_T^* also increases for firms at the exit cutoff. This is because

$$A_1 > A_2.$$

Third, as

$$T = log(x_T)$$

and employment equals $x_T - 1$, both the number of layers and employment increase for firms at the exit cutoff.

Third, I prove the result that the firm size distribution and the distribution of the number of layers move to the right in the FOSD sense when MT improves. Similar to the proof in the paper, the key thing to show is that for any $c > 1$

$$S(c\bar{\theta}_2, A_2) > S(c\bar{\theta}_1, A_1),$$

$$x_T(c\bar{\theta}_2, A_2) > x_T(c\bar{\theta}_1, A_1),$$

$$l(c\bar{\theta}_2, A_2) > l(c\bar{\theta}_1, A_1)$$

and

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

First, note that

$$\frac{S(c\bar{\theta}, A)}{S(\bar{\theta}, A)} = c$$

and

$$\frac{x_T(c\bar{\theta}, A)}{x_T(\bar{\theta}, A)} = c,$$

which do not depend on A and $\bar{\theta}$. As it has already been shown that $S(\bar{\theta}_2, A_2) > S(\bar{\theta}_1, A_1)$ and $x_T(\bar{\theta}_2, A_2) > x_T(\bar{\theta}_1, A_1)$, I conclude that

$$S(c\bar{\theta}_2, A_2) > S(c\bar{\theta}_1, A_1)$$

and

$$x_T(c\bar{\theta}_2, A_2) > x_T(c\bar{\theta}_1, A_1),$$

which mean that the distributions of revenue and output move to the right in the FOSD sense when MT improves. Second, as $l(\theta, A) = x_T(\theta, A) - 1$ and $T(\theta, A) = \log(x_T(\theta, A))$ for any (θ, A), I also have

$$l(c\bar{\theta}_2, A_2) > l(c\bar{\theta}_1, A_1)$$

and

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

which mean that the distributions of employment and the number of layers move to the right in the FOSD sense as well when MT improves. QED.

The intuitions behind this proposition are the same as the intuitions in the paper. The management shock common across all firms benefits firms with more layers disproportionately more. As a result, the least efficient firms exit the market and the most efficient firms thrive. Therefore, the firm size distribution and the distribution of the number of layers move to the right in the FOSD sense after the management shock.

Fourth, I discuss the welfare implication of an improvement in MT. In order to derive results on welfare, I solve for the exit cutoff and the adjusted market size first. Based on the discussion in the previous section, I derive the ZCP condition and the FE condition using equation (84). More specifically, the ZCP condition states that

$$(1-\beta)A^{\sigma}\bar{\theta}\Big(\frac{\beta}{ge}\Big)^{\sigma-1} = f - ge. \tag{86}$$

The FE condition says that

$$\int_{\theta=\bar{\theta}}^{\infty}(\pi(\theta, A) - f)g(\theta)d\theta = f_e,$$

which can be further reduced to

$$(f - ge)\frac{(\theta_{min}/\bar{\theta})^k}{(k-1)} = f_e.$$

Therefore, the exit cutoff can be solved as

$$\bar{\theta} = \theta_{min}\Big(\frac{(f-ge)}{f_e(k-1)}\Big)^{\frac{1}{k}}. \tag{87}$$

Substituting equation (32) into equation (29) leads to the solution of the adjusted market size A as

$$A^{\sigma} = \frac{(f-ge)(ge)^{\sigma-1}}{(1-\beta)\beta^{\sigma-1}\bar{\theta}} = \frac{(f-ge)^{1-\frac{1}{k}}(ge)^{\sigma-1}((k-1)f_e)^{\frac{1}{k}}}{(1-\beta)\beta^{\sigma-1}\theta_{min}}. \tag{88}$$

Now, I derive an explicit expression of welfare in the closed economy. Since welfare equals p_h (i.e., wage offered in the homogeneous sector), I calculate the welfare using the following equation:

$$\Big[WCper(A,\bar{\theta}) - \psi LDper(A,\bar{\theta})\Big]P_1(\bar{\theta},A)^{\sigma-1} + \frac{(1-\gamma)A^{\sigma}}{\gamma} = p_h^{1-\frac{(1-\gamma)(\sigma-1)}{\gamma}}L. \tag{89}$$

Based on equations (82), (83), (87) and (88), I obtain that[57]

$$P_1(\bar{\theta},A)^{\sigma-1} = \Big(\frac{ge}{\beta}\Big)^{\sigma-1}\frac{(k-1)}{k\bar{\theta}}, \tag{90}$$

$$LDper(A,\bar{\theta}) = \frac{k}{k-1}x_T(\bar{\theta}) - 1, \tag{91}$$

and

$$WCper(A,\bar{\theta}) = LDper(A,\bar{\theta})ge = ge\Big[\frac{k}{k-1}x_T(\bar{\theta}) - 1\Big], \tag{92}$$

where $x_T(\bar{\theta}) = \frac{(\sigma-1)(f-ge)}{ge}$ is the output level of the smallest firms in equilibrium in terms of the final composite good. Note that both $LDper(A,\bar{\theta})$ and $WCper(A,\bar{\theta})$ must be greater than or equal to zero. Thus, it must be true that

$$x_T(\bar{\theta}) \geq \frac{k-1}{k},$$

which implies that[58]

$$f > \frac{\sigma - \frac{1}{k}}{\sigma - 1}b\psi e. \tag{93}$$

After having substituting equations (90) to (92) into the L.H.S. of equation (89), I end up with

$$\begin{aligned} & \psi(be-1)\Big[\frac{k(\sigma-1)(f-ge)}{(k-1)ge} - 1\Big]\Big(\frac{ge}{\beta}\Big)^{\sigma-1}\frac{(k-1)}{k\theta_{min}}\Big(\frac{f_e(k-1)}{(f-ge)}\Big)^{\frac{1}{k}} \\ + \quad & \frac{(1-\gamma)}{\gamma}\frac{(f-ge)^{1-\frac{1}{k}}(ge)^{\sigma-1}((k-1)f_e)^{\frac{1}{k}}}{(1-\beta)\beta^{\sigma-1}\theta_{min}}. \end{aligned}$$

Denote that

$$B(b) \equiv \frac{(ge)^{\sigma-1}(ge-\psi)}{(f-ge)^{\frac{1}{k}}}\Big[\frac{(\sigma-1)(f-ge)}{ge} - \frac{k-1}{k}\Big] + \frac{(1-\gamma)\sigma}{\gamma}(f-ge)^{1-\frac{1}{k}}(ge)^{\sigma-1}. \tag{94}$$

Therefore, welfare in the closed economy is

$$p_h^c = \Big[\frac{(f_e(k-1))^{\frac{1}{k}}B(b)}{\beta^{\sigma-1}\theta_{min}L}\Big]^{\frac{\gamma}{\gamma-(1-\gamma)(\sigma-1)}}, \tag{95}$$

[57]Note that price charged by a firm with a efficiency level θ is $A\theta^{q/\sigma}x_T(\theta)^{-1/\sigma} = \frac{ge}{\beta}$. This means that all firms charge the same price, as the quality of their goods differs.

[58]This condition must be true, since the entrepreneur can always produce one unit of good by herself. This implies that output level of any active firm must be greater than or equal to one, if all firms use management hierarchies to produce in equilibrium. As in the main context of the paper, I focus on the case in which there are no self-employed entrepreneurs.

and there are welfare gains from an improvement in MT if and only if one of the following conditions holds.

$$\frac{d\ln B(b)}{db} > 0, \quad \frac{\sigma-1}{\sigma} > \gamma$$

or

$$\frac{d\ln B(b)}{db} < 0, \quad \frac{\sigma-1}{\sigma} < \gamma.$$

The above condition is not intuitive, since it involves many parameters. Now, I derive a necessary and sufficient condition for the existence of welfare gains from better MT. From equation (94), I conclude that $B^{'}(b) > 0$ if both

$$(f - b\psi e)^{1-\frac{1}{k}}(b\psi e)^{\sigma-1}$$

and

$$(b\psi e)^{\sigma-2}(f - b\psi e)^{1-\frac{1}{k}}(b\psi e - \psi)$$

increase in t. Calculation shows that the first expression increases with b, since

$$\frac{\sigma-1}{b\psi e} > \frac{1-\frac{1}{k}}{f - b\psi e}$$

implied by condition (93). Moreover, the second expression also increases with b, since

$$\frac{\sigma-2}{b\psi e} + \frac{1}{b\psi e - \psi} > \frac{\sigma-1}{b\psi e} > \frac{1-\frac{1}{k}}{f - b\psi e}.$$

Therefore, $B(b)$ increases with b. As a result, welfare increases after an improvement in MT *if and only if*

$$\frac{\sigma-1}{\sigma} > \gamma.$$

Finally, I discuss the condition under which the type of equilibrium I investigate exists. Namely, when does every labor submarket have unemployment? Every *employed* worker receives $be\psi$ as wage compensation in the CES sector and every worker receive p_h as wage compensation in the homogeneous sector. Therefore, a necessary and sufficient condition for the equilibrium to exist is that

$$p_h^c \leq \psi(be - 1).$$

Note that the above condition contains only exogenous parameters, as p_h^c only contains exogenous parameters. Furthermore, this condition puts an upper bound on L when $\frac{\sigma-1}{\sigma} > \gamma$ and a lower bound on L when $\frac{\sigma-1}{\sigma} < \gamma$.

9.5 A Continuous Number of Layers: the Open Economy

9.5.1 Equilibrium Outcomes

I calculate the WGT using following several steps. First, optimal allocation of output between two symmetric markets implies that

$$q_d(\theta) = \frac{q\tau^{\sigma-1}}{1 + \tau^{\sigma-1}} \tag{96}$$

and

$$p_x(\theta) = \tau p_d(\theta) = \frac{\tau ge}{\beta}, \tag{97}$$

where $p_x(\theta)$ and $p_d(\theta)$ are prices charged in the foreign market and in the domestic market respectively, if the firm exports.

Second, I solve for the product market equilibrium. First, the indifference condition between exporting or not is

$$\tau^{\sigma-1} f_x = \frac{A^\sigma}{\sigma} \bar{\theta}_x \Big(\frac{\beta}{ge}\Big)^{\sigma-1}. \tag{98}$$

Second, the FE condition in the open economy becomes

$$\begin{aligned} &\int_{\theta=\bar{\theta}}^{\bar{\theta}_x} (f - ge) \frac{\theta}{\bar{\theta}} g(\theta) d\theta + \int_{\theta=\bar{\theta}_x}^{\infty} (f - ge) \frac{\bar{\theta}_x}{\bar{\theta}} \Big(1 + \frac{1}{\tau^{\sigma-1}}\Big) \frac{\theta}{\bar{\theta}_x} g(\theta) d\theta \\ &-(f - ge)\Big(\frac{\theta_{min}}{\bar{\theta}}\Big)^k - f_x \Big(\frac{\theta_{min}}{\bar{\theta}_x}\Big)^k = f_e, \end{aligned} \tag{99}$$

Using equations (98) to (99) and the ZCP condition in equation (37), I solve for the exit cutoff which equals

$$\bar{\theta} = \theta_{min} \Big(\frac{(f - ge)}{f_e(k-1)}\Big)^{\frac{1}{k}} \left[1 + \Big(\frac{f_x \tau^{\sigma-1}}{f - ge}\Big)^{-k+1} \frac{1}{\tau^{\sigma-1}}\right]^{\frac{1}{k}}. \tag{100}$$

And, the exporting cutoff is

$$\bar{\theta}_x = \bar{\theta} \frac{f_x \tau^{\sigma-1}}{(f - ge)}. \tag{101}$$

Substituting equation (101) into equation (98) yields the solution for the adjusted market size which is equal to

$$A^\sigma = \frac{(f - ge)(ge)^{\sigma-1}}{(1-\beta)\beta^{\sigma-1}\bar{\theta}} = \frac{(f - ge)^{1-\frac{1}{k}}(ge)^{\sigma-1}((k-1)f_e)^{\frac{1}{k}}}{(1-\beta)\beta^{\sigma-1}\theta_{min}} \left[1 + \Big(\frac{f_x \tau^{\sigma-1}}{f - ge}\Big)^{-k+1} \frac{1}{\tau^{\sigma-1}}\right]^{-\frac{1}{k}}. \tag{102}$$

Third, I solve for the welfare (i.e., the workers' expected payoff). Using the labor-market-clearing conditions in equations (39) and (40), I restate the labor-market-clearing condition as

$$\begin{aligned} &\psi(be-1)\left[\frac{k}{k-1}\frac{(\sigma-1)(f-ge)}{ge}\Big[1 + \tau^{1-\sigma}\Big(\frac{f_x\tau^{\sigma-1}}{(f-ge)}\Big)^{-k+1}\Big] - 1\right]\Big(\frac{ge}{\beta}\Big)^{\sigma-1} \\ &\frac{(k-1)}{k\theta_{min}}\Big(\frac{f_e(k-1)}{(f-ge)}\Big)^{\frac{1}{k}}\left[1 + \Big(\frac{f_x\tau^{\sigma-1}}{f-ge}\Big)^{-k+1}\frac{1}{\tau^{\sigma-1}}\right]^{-\frac{1}{k}}\Big[1 + \tau^{1-\sigma}\Big(\frac{f_x\tau^{\sigma-1}}{(f-ge)}\Big)^{-k+1}\Big]^{-1} \\ &+\frac{(1-\gamma)}{\gamma}\frac{(f-ge)^{1-\frac{1}{k}}(ge)^{\sigma-1}((k-1)f_e)^{\frac{1}{k}}}{(1-\beta)\beta^{\sigma-1}\theta_{min}}\left[1 + \Big(\frac{f_x\tau^{\sigma-1}}{f-ge}\Big)^{-k+1}\frac{1}{\tau^{\sigma-1}}\right]^{-\frac{1}{k}} \\ = \; & p_h^{1-\frac{(1-\gamma)(\sigma-1)}{\gamma}} L, \end{aligned} \tag{103}$$

where the first part of the left hand side (LHS) of the above equation is related to the number of job applicants in the CES sector, while the second part of the LHS of the above equation is related to the number of workers in the homogeneous sector. In total, it states that the number of workers who seek jobs is equal to the fixed labor endowment, L. Solving equation (103), I obtain the welfare in the open economy as

$$p_h^o = \left[\frac{(f_e(k-1))^{\frac{1}{k}} A(b,\tau)}{\beta^{\sigma-1}\theta_{min} L}\right]^{\frac{\gamma}{\gamma-(1-\gamma)(\sigma-1)}}, \tag{104}$$

where

$$\begin{aligned} A(b,\tau) \equiv \quad & \psi(be-1)\left[\frac{k}{k-1}\frac{(\sigma-1)(f-ge)}{ge}\left[1+\tau^{1-\sigma}\left(\frac{f_x\tau^{\sigma-1}}{(f-ge)}\right)^{-k+1}\right]-1\right](ge)^{\sigma-1} \\ & \frac{(k-1)}{k}\left(\frac{1}{(f-ge)}\right)^{\frac{1}{k}}\left[1+\left(\frac{f_x\tau^{\sigma-1}}{f-ge}\right)^{-k+1}\frac{1}{\tau^{\sigma-1}}\right]^{-\frac{1}{k}}\left[1+\tau^{1-\sigma}\left(\frac{f_x\tau^{\sigma-1}}{(f-ge)}\right)^{-k+1}\right]^{-1} \\ & +\frac{(1-\gamma)\sigma}{\gamma}(f-ge)^{1-\frac{1}{k}}(ge)^{\sigma-1}\left[1+\left(\frac{f_x\tau^{\sigma-1}}{f-ge}\right)^{-k+1}\frac{1}{\tau^{\sigma-1}}\right]^{-\frac{1}{k}}. \end{aligned} \tag{105}$$

Therefore, the change in welfare is

$$WGT(b,\tau) \equiv \frac{p_h^o}{p_h^a} - 1 = \left[\frac{A(b,\tau)}{B(b)}\right]^{\frac{\gamma}{\gamma-(1-\gamma)(\sigma-1)}} - 1. \tag{106}$$

Note that WGT are not guaranteed.[59] However, I am going to derive some sufficient conditions assuring WGT in next subsection.

Before proceeding to the discussion of WGT, I discuss the condition under which the type of equilibrium I investigate exists. The equilibrium exists and is unique if

$$p_h^o \leq \psi(be-1),$$

where p_h^o is defined in equation (103). Note that the above condition only contains exogenous parameters (i.e., primitives of the model). Furthermore, this condition puts an upper bound on L when $\frac{\sigma-1}{\sigma} > \gamma$ and a lower bound on L when $\frac{\sigma-1}{\sigma} < \gamma$.

9.5.2 Management Quality, the Trade Share, and the Welfare Gains from Trade

In this subsection, I show how management quality affects the trade share and discuss how it affects the WGT. I discuss the relationship between management quality and the trade share first. Let λ be the share of domestic consumption in total expenditure on the CES goods. Using equations (96)-(97) and (100)-(102), I derive the domestic consumption share as

$$\lambda(\tau,b) = \frac{\tau^{\sigma-1}\left(\frac{\bar{\theta}_x}{\bar{\theta}}\right)^{k-1}}{1+\tau^{\sigma-1}\left(\frac{\bar{\theta}_x}{\bar{\theta}}\right)^{k-1}} \tag{107}$$

[59] I use $WGT(b,\tau)$ to denote the welfare *change* from opening up to trade to save notations. For some parameter values, $WGT(b,\tau)$ might be negative, which implies welfare losses from trade.

and the import share as

$$1-\lambda(\tau,b)=\frac{1}{1+\tau^{\sigma-1}\left(\frac{\bar{\theta}_x}{\theta}\right)^{k-1}}. \tag{108}$$

Since $\frac{\bar{\theta}_x}{\theta}$ increases in b from equation (101), the better the MT, the bigger the share of exporting firms. Furthermore, equation (108) shows that the better the MT, the bigger the trade share.

Next, I derive sufficient and necessary conditions under which there are WGT. Before moving on, I make the following simplifying notations.

$$t\equiv ge=b\psi e;$$

$$C(t,\tau)\equiv\frac{f_x^{-k+1}\tau^{-k(\sigma-1)}}{(f-ge)^{-k+1}}.$$

Note that $C(t,\tau,f_x)$ decreases with t, τ and f_x. Calculation shows that

$$\frac{dA(b,\tau)}{d\tau}<0,$$

if and only if

$$Q(\sigma,\gamma,f_x,f,\tau,\psi)=\frac{(\sigma-1)(f-t)}{t}+\frac{(1-\gamma)\sigma}{\gamma}\frac{f-t}{t-\psi}-\frac{k^2-1}{k(1+C(t,\tau))}>0. \tag{109}$$

Therefore, there are WGT (i.e., $WGT(b,\tau)>0$), if and only if one of the following two conditions holds:

$$Q(\sigma,\gamma,f_x,f,\tau,\psi)>0,\ \ \frac{\sigma-1}{\sigma}>\gamma$$

or

$$Q(\sigma,\gamma,f_x,f,\tau,\psi)<0,\ \ \frac{\sigma-1}{\sigma}<\gamma.$$

I will discuss the economic interpretation of these conditions in what follows.

Since the above conditions are not intuitive, I derive sufficient conditions that guarantees WGT. I focus on the case in which $\frac{\sigma-1}{\sigma}>\gamma$ first. In this case, the inequality in equation (109) holds for all possible values of k and τ if

$$f>\Big[1+\frac{(\sigma-1)k}{\sigma^2-\sigma+1}\Big]t=\Big[1+\frac{(\sigma-1)k}{\sigma^2-\sigma+1}\Big]\psi eb. \tag{110}$$

The above condition indicates that when the elasticity of substitution is big, there are WGT if the management quality is high. Note that the condition in equation (110) is a sufficient condition for the existence of WGT for *all* possible levels of the iceberg trade cost. In other words, there might be welfare losses for certain reductions in the iceberg trade cost, if the condition in equation (110) is not satisfied.

In the case in which $\frac{\sigma-1}{\sigma}\leq\gamma$, a sufficient condition for the existence of the WGT is

$$\psi eb\Big[1+\frac{k^2-1}{2k}\frac{\sigma-1}{\sigma^2-\sigma+1}\Big]>\frac{k^2-1}{2k}\frac{\sigma-1}{\sigma^2-\sigma+1}\psi+f. \tag{111}$$

The above condition says that when the elasticity of substitution is small, there are WGT if the management quality of firms in the CES sector is low. The economic reasoning follows the same logic that I have discussed above. In summary, WGT are not guaranteed in a world with multiple frictions. In particular, whether or not there are WGT crucially depends on whether or not resources that are reallocated after the trade liberalization are used inefficiently before the trade liberalization.

Now, I show that under one of the above sufficient conditions (i.e., equation (110) and $\frac{\sigma-1}{\sigma} > \gamma$), the complementarity result holds. Namely, the better is the MT, the larger are the WGT. I focus on this condition, since the parameters of the calibrated model presented in the next section satisfy this condition. Recall that the WGT are

$$WGT(b,\tau) \equiv \frac{p_h^o}{p_h^a} - 1 = \Big[\frac{B(b)}{A(b,\tau)}\Big]^{\frac{\gamma}{(1-\gamma)(\sigma-1)-\gamma}} - 1.$$

Denote

$$D(b,\tau) \equiv \frac{B(b)}{A(b,\tau)}.$$

As long as $\frac{d\ln D(b)}{db} < 0$, the complementarity result holds when $\frac{\sigma-1}{\sigma} > \gamma$. Calculation shows that

$$\frac{d\ln D(b)}{db} < 0,$$

since $t^2 + \psi(f-2t) > 0$, $C(t,\tau) > 0$ and

$$f > \Big[1 + \frac{(\sigma-1)k}{\sigma^2-\sigma+1}\Big]t.$$

Therefore, there are the WGT, and better MT makes the WGT larger if

$$f > \Big[1 + \frac{(\sigma-1)k}{\sigma^2-\sigma+1}\Big]t, \quad \frac{\sigma-1}{\sigma} > \gamma.$$

9.5.3 A Comparison to the ACR Formula

I derive the formula for the WGT in this subsection. Among the two aggregate statistics appearing in the ACR formula, one statistic (i.e., the domestic consumption share) is derived in equation (107). The other one which is the elasticity of the trade share with respect to the variable trade cost is calculated as

$$\epsilon = \frac{\partial \ln\left(1-\lambda(\tau,b)\right)/\lambda(\tau,b)}{\partial \ln \tau} = -(\sigma-1)k.$$

Calculation shows that the WGT are

$$WGT(b,\tau) = \lambda(\tau,b)^{\frac{\gamma}{k(\gamma-(1-\gamma)(\sigma-1))}} \left[\frac{\Big[1 + \frac{(1-\gamma)\sigma}{\gamma(\sigma-1)}\frac{be}{(be-1)}\Big]x_T(\bar{\theta}) - \frac{k-1}{k}\lambda(\tau,b)}{\Big[1 + \frac{(1-\gamma)\sigma}{\gamma(\sigma-1)}\frac{be}{(be-1)}\Big]x_T(\bar{\theta}) - \frac{k-1}{k}}\right]^{\frac{\gamma}{\gamma-(1-\gamma)(\sigma-1)}}, \tag{112}$$

where

$$x_T(\bar{\theta}) = \frac{(\sigma-1)(f-t)}{t} \tag{113}$$

is the output level of the smallest firms in equilibrium. Note that this variable is unit-free, since all variables in the model are denominated in terms of the final composite good. Equation (113) is a function of ratios as well, since f/t is the ratio of the fixed production cost to the equilibrium wage offered in the CES sector. Thus, the value of equation (113) does not depend on the units that are used to measure the output in equilibrium.

I need information on the expenditure share on the CES goods (i.e., γ), the management quality (i.e., $\frac{1}{b}$) and the output level of the smallest firms (i.e., $x_T(\bar{\theta})$) in order to evaluate the welfare change due to opening up to trade. In an extreme case in which there is no outside sector, the above equation is simplified to

$$WGT(b,\tau) = \lambda(\tau,b)^{\frac{1}{k}} \frac{x_T(\bar{\theta}) - \frac{k-1}{k}\lambda(\tau,b)}{x_T(\bar{\theta}) - \frac{k-1}{k}}.$$

Even in this extreme case, information on three variables is still needed to evaluate the welfare change from opening up to trade. A new variable that does not show up in the ACR formula is the output level of the smallest firms in equilibrium. In summary, in a world with information frictions which validate the use of management hierarchies, evaluating the WGT requires more information than that contained in the ACR formula.

9.6 Another Type of Equilibrium

The paper considers one type of equilibrium in which there is unemployment in every labor submarket. This type of equilibrium must uniquely exist under restrictions on parameter values. However, when the outside option of workers is high enough in equilibrium, firms in the CES sector that pay the lowest wages are forced to raise their wages up to the level of the sum of the workers' outside option and the cost to exert effort. As a result, job applicants who enter the labor submarkets of these firms are fully employed. In other words, these labor submarkets are perfectly competitive markets. Therefore, this type of equilibrium contains both labor submarkets that have no unemployment and labor submarkets that have unemployment. This is the third type of equilibrium which I have not considered in the paper.[60]

I change the setup of the model slightly in order to simplify the analysis. Different from the paper, I assume that there is only one sector (i.e., the CES sector) in the economy (i.e., $\gamma = 1$), and there are infinitely many workers who have a fixed outside option p_h (the net payoff they receive from working elsewhere) and can choose whether or not to enter the CES sector. Admittedly, I can not discuss welfare implications under this alternative specification of the model, since the welfare of workers is fixed by an exogenous parameter (i.e., p_h). However, I show that qualitative results of the model except for the results on welfare hold in this type of equilibrium as well using a numerical example.

There are four equilibrium conditions in total under the above alternative specifications. First, the two equilibrium conditions for the product market (i.e., the CES sector)

[60]The first type of equilibrium is an equilibrium in which every labor submarket is a competitive market that has a uniform wage and no unemployment. This equilibrium exists when the cost of exerting effort were zero (i.e., no incentive problems). The second type of equilibrium is an equilibrium in which every labor submarket has unemployment. This type of equilibrium exists when the cost of exerting effort is high enough.

are still the ZCP condition and the FE condition (i.e., equations (23) and (24)). Second, the outside option of workers (i.e., p_h) pins down the number of job applicants in each labor submarket (e.g., equation (26)) and the total number of job applicants in the CES sector (e.g., equation (28)). Finally, the aggregate income E is still determined by the market-clearing condition of the final composite good that is equation (30). Note that as the pool of potential entrepreneurs is large enough, the FE condition holds with equality. The indifference condition of workers holds with equality as well, since there are infinitely many workers who are willing to enter the CES sector, as long as the expected payoff obtained from entering the CES sector is bigger than or equal to p_h. In total, I end up with four equations (i.e., equations (23), (24), (28) and (30)) and four endogenous variables: the exit cutoff $\bar{\theta}$, the mass of firms M, the aggregate labor demand of the CES sector L_c and the aggregate income E. I eliminate one condition (i.e., equation (30)) and normalize the ideal price index of the CES goods to one.

I show that qualitative results derived in the paper hold in the current case as well. As it is hard to prove results analytically, I simulate the model to show the existence of the pro-competitive effect of improvements in MT. The approach to solve the model in the current case is similar to the simulation algorithm I have used in the paper. More specifically, first I assume that the outside option of workers (i.e., p_h) does no bind when firms choose their optimal wage schedules (i.e., assume that $w_i(\theta) - \psi \geq p_h \; \forall \; (i, \theta)$), and then solve for the value of $(\bar{\theta}, A)$. Next, I check whether or not $w_i(\theta) - \psi \geq p_h$ for all (i, θ), and I consider the case in which $w_i(\theta) - \psi < p_h$ for some (i, θ).[61] In the simulation example, I find that among firms that have two layers (i.e., $T = 1$), the smallest ones would have offered wages lower than $p_h + \psi$, if they ignored the outside option of workers. Thus, the optimization problem of these firms has to be resolved, and the outside option of workers has to be taken into account when the firm designs the optimal wage schedule.

I solve the optimization problem of firms that are *constrained* to offer wage schedules in two steps. First, I derive the cutoff for θ below which firms have to raise their wages up to $p_h + \psi$. As only firms that have two layers may be constrained to offer the wage schedules in the simulated example, I only consider firms that have two layers now. Without being constrained by the outside option of workers, a firm that has two layers would offer wage $w_i(\theta)$ at

$$w_i(\theta) = b\psi\Big(\frac{A\beta\theta^{\frac{1}{\sigma}}}{2b\psi}\Big)^{\frac{\sigma}{\sigma+1}}.$$

Therefore, the cutoff θ_1 is

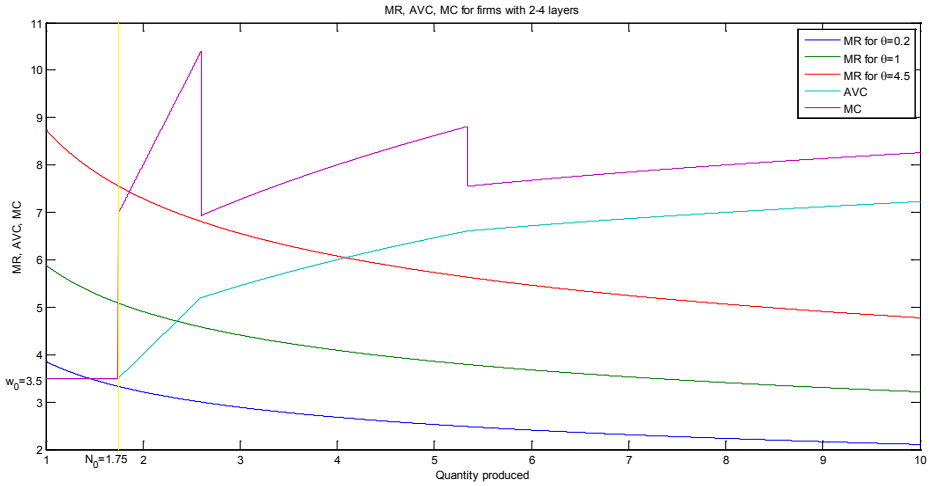
$$\theta_1 = \Big(\frac{2}{A\beta}\Big)^{\sigma}\frac{(p_h+\psi)^{\sigma+1}}{b\psi}. \tag{114}$$

Next, I discuss the optimization problem for a firm that has the demand draw below θ_1. Its objective function can be stated as

$$\begin{aligned} \max_{w,N} \quad & A\theta^{\frac{1}{\sigma}}N^{\beta} - wN \\ s.t. \quad & w \geq p_h + \psi, \\ & \frac{w}{bN} \geq \psi, \end{aligned}$$

[61] As workers' wage at layer i does not change monotonically with firm size, it is not necessarily true that $w_i(\theta) - \psi < p_h$ for the smallest firms.

Figure 11: The Modified AVC Curve and MC curve



where the first constraint is the lower bound on wage that the firm can offer, and the second constraint is the incentive compatibility (IC) constraint for the production workers to exert effort.[62] There are two types of solutions to the above optimization problem depending on the value of $\bar{\theta}$. First, if $\bar{\theta} \geq \frac{\theta_1}{2^\sigma}$, the solution is

$$w^* = p_h + \psi, \quad N^* = \frac{p_h + \psi}{b\psi} \tag{115}$$

for $\theta \in [\bar{\theta}, \theta_1]$. Second, if $\bar{\theta} < \frac{\theta_1}{2^\sigma}$, the solution for firms with $\theta \in [\frac{\theta_1}{2^\sigma}, \theta_1]$ is still characterized by equation (115). The solution for firms with $\theta \in [\bar{\theta}, \frac{\theta_1}{2^\sigma}]$ is characterized by

$$w^* = p_h + \psi, \quad N^*(\theta) = \left(\frac{A\beta\theta^{\frac{1}{\sigma}}}{p_h + \psi} \right)^\sigma. \tag{116}$$

The average variable cost (AVC) and the marginal cost (MC) are slightly different in the current case compared with the case discussed in paper. First, the AVC curve given $T = 1$ becomes a horizontal line at $w_0 \equiv p_h + \psi$ for output less than $N_0 \equiv \frac{p_h+\psi}{b\psi}$. Firms cannot reduce wages paid to production workers below $p_h + \psi$, even if output and the span of control fall below $\frac{p_h+\psi}{b\psi}$. Second, the MC curve also becomes a horizontal line at $w_0 = p_h + \psi$ for output less than N_0, as firms pay $p_h + \psi$ to every production worker for any output level smaller than $\frac{p_h+\psi}{b\psi}$. This implies that there is a discontinuous increase in the MC when output exceeds N_0. When output exceeds N_0, a firm has to increase wage payment to *all* existing workers in order to produce more, while wage paid to every worker is a constant (i.e., w_0) for any output level less than N_0. Figure 11 shows the AVC and MC curves of the firm in the current case.

Due to these changes in the cost functions, there are three cases to consider when the firm produces output using a management hierarchy with two layers. First, if the demand draw is bigger than θ_1, which implies that the optimal output exceeds N_0, output

[62] Note that the span of control is $\frac{1}{N}$ for the entrepreneur.

and wage are the same as the ones I have derived in the paper (i.e., equation (14) in the paper and the equation for $w_i(\theta)$ on page 48 of the paper). This case is illustrated by the marginal revenue (MR) curve for $\theta = 4.5$ in Figure 11. Second, if the demand draw of a firm is between $\frac{\theta_1}{2^\sigma}$ and θ_1, the firm produces output at N_0, as there is a discontinuous increase in the MC at N_0. This case is represented by the MR curve for $\theta = 1$ in Figure 11. Finally, if the demand draw of a firm is smaller than $\frac{\theta_1}{2^\sigma}$, the firm faces a constant MC. The optimal wage and output are solved in equation (116). Note that whether or not the third case appears in equilibrium depends on whether or not $\bar{\theta}$ is smaller than $\frac{\theta_1}{2^\sigma}$. In the numerical example I am going to show, only the first two cases appear in equilibrium.

Simulation results show that qualitative results of the paper still hold in the current case. First, the exit cutoff for demand draw θ increases when MT improves, which is shown in Table 10. Second, Figure 12 shows that the firm size distribution moves to the right in the FOSD sense when the quality of MT improves. Third, when MT improves, aggregate productivity of active firms increases due to both the within-firm effect and the between-firm effect. Finally, all surviving firms increase the number of layers weakly when the quality of MT improves. The last two results are shown in Figure 13.

Table 10: The Cutoffs

	$\bar{\theta}$ (the exit cutoff)	θ_1
b=1.45	4.4498	4.5998
b=1.38	4.6991	5.4608

$\psi = 0.44$, $p_h = 0.56$, $\sigma = 3.8$, $k = 1.1$, $\theta_{min} = 1$

The are two effects that make the exit cutoff increase when MT improves. First, an improvement in MT benefits firms having more layers disproportionately more when they are *not* constrained to offer the wage schedules (i.e., firms having demand draw $\theta(> \theta_1)$). This effect is due to the cost structure of the firm that have been discussed in the paper. Second, firms with the lowest demand draws among firms that have two layers (i.e., firms with demand draw $\theta \in [4.6991, 5.4608]$ in the simulation) can not fully realize the benefit of improved MT. They offer the lowest wages to production workers before MT improves. When MT improves, they are forced to set the wage at $p_h + \psi$ which is higher than what they would choose without the constraint of the workers' outside option (i.e., p_h). The second effect reinforces the first one, as it also favors firms that have better demand draws. In total, these two effects together make the exit cutoff increase and intensify market competition.

Changes in firm-level and aggregate-level outcomes are results of the intensified competition triggered by an improvement in MT. First, surviving firms become bigger on average when MT improves, since the smallest firms exit the market and the biggest firm expand. Second, aggregate productivity increases, as the least productive firms exit the market and the productivity of surviving firms increases. Finally, surviving firms weakly increase the number of layer, as better MT reduces their labor costs and incentives firms to produce more. In sum, the qualitative results of an improvement in MT discussed in the paper hold in the current case as well.

The takeaway from this subsection is that the key economic force of an improvement in MT discussed in the paper (i.e., firms with better demand draws benefit disproportionately more from such an improvement) also exists in the case in which some labor

Figure 12: The Firm Size Distribution

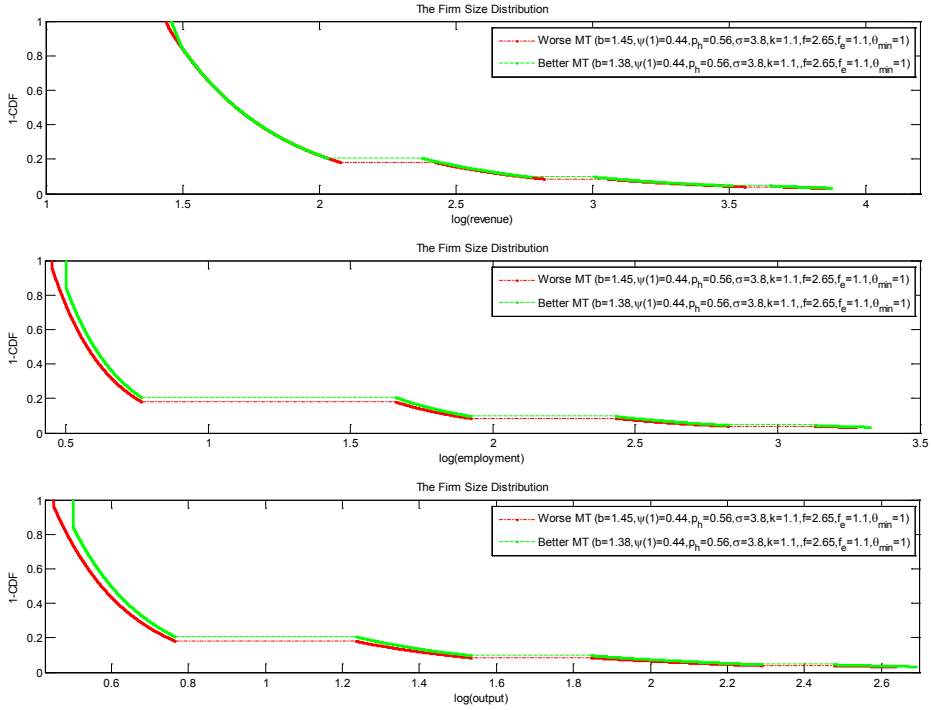
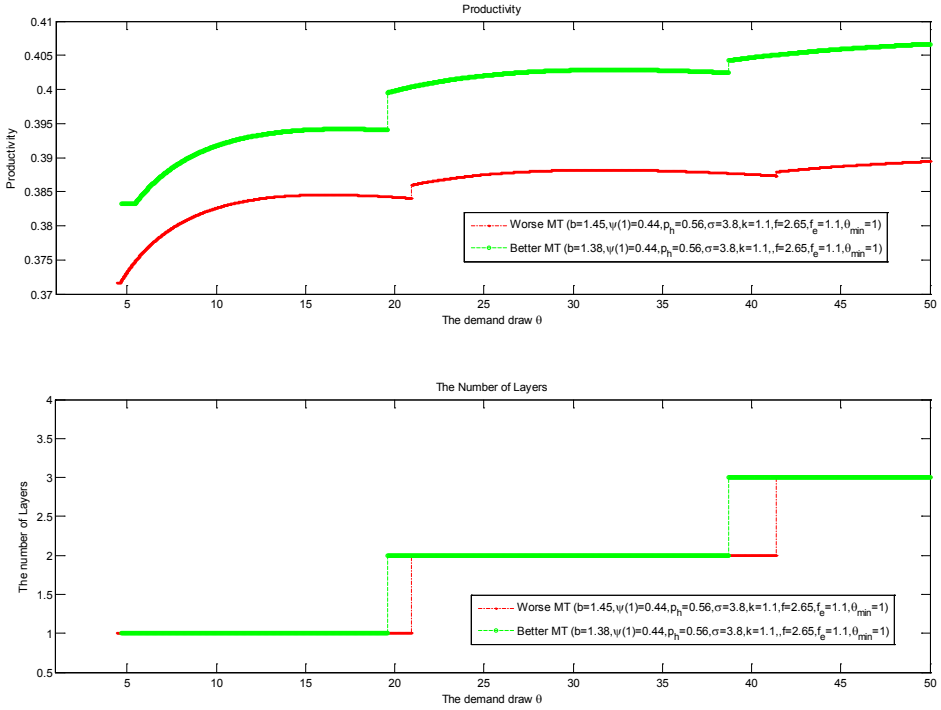


Figure 13: Aggregate Productivity and the Number of Layers



submarkets don't have unemployment. Moreover, there is another channel through which this uneven effect works as what I have discussed above.

References:

1. Akcigit, Ufuk, Harun Alp, and Michael Peters (2014): "Lack of Selection and Imperfect Managerial Contracts: Firm Dynamics in Developing Countries," (Mimeo, Yale University).

2. Arkolakis, Costas, Arnaud Costinot, and Andres Rodriguez-Clare (2012): "New Trade Models, Same Old Gains," *American Economic Review* 102: 94-130.

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

4. 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.

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

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

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

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

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

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

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

12. 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.

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

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

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

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

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

18. Chen, Cheng (2014): "Management Quality and Firm Organization and International Trade," `http://scholar.princeton.edu/sites/default/files/ccfour/files/management_firm_organization_and_international_trade_0.pdf` (Mimeo, Princeton University).

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

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

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

22. Eaton, Jonathan, and Samuel Kortum (2002): "Technology, Geography, and Trade," *Econometrica* 70(5): 1741-1779.

23. 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.

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

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

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

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

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

29. Ghironi, Fabio, and Marc J. Melitz (2005): "International Trade and Macroeconomic Dynamics with Heterogeneous Firms," *Quarterly Journal of Economics* 120: 865-915.

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

31. 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.

32. Hall, Robert (2006): "Sources and Mechanisms of Cyclical Fluctuations in the Labor Market," (Mimeo, Stanford University).

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

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

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

36. Hsieh, Chiang-Tai, and Benjamin A. Olken (2014): "The Missing "Missing Middle"," NBER Working Paper 19966.

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

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

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

40. Krugman, Paul. R. (1980): "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review* 70: 950-959.

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

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

43. Melitz, Marc J., and Stephen J, Redding (2013): "New Trade Models, New Welfare Implications," NBER Working Paper 18919.

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

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

46. Powell, Michael (2013): "Productivity and Credibility in Industry Equilibrium," (Mimeo, Northwestern University).

47. Poschke, Markus (2014): "The firm size distribution across countries and skill-biased change in entrepreneurial technology," `http://markus-poschke.research.mcgill.ca/papers/mposchke_skillbias.pdf` (Mimeo, McGill University).

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

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

50. 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.

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

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