

Management Quality and Firm Hierarchy in Industry Equilibrium*

Cheng Chen[†]

Abstract. I incorporate a monitoring-based firm hierarchy into an industry equilibrium model with heterogeneous firms. I then use the theory to study aggregate impact of an economy-wide improvement in monitoring efficiency. This shock generates a *selection effect* which favors more hierarchical (i.e., more layers) firms. Interestingly, these implications depend on firms' *heterogeneous* choices about their hierarchy and completely disappear when firms are homogeneous in terms of the number of layers inside the hierarchy.

Keywords. management quality, monitoring and incentives; firm hierarchy; resource allocation

JEL Classification. D21; D23; L22; L23

*I am grateful to Gene Grossman, Stephen Redding, and Esteban Rossi-Hansberg for invaluable guidance. I thank Pol Antràs, Nick Bloom, Davin Chor, Liang Dai, Taiji Furusawa, Jan Grobovsek, Maria Guadalupe, Jota Ishikawa, Hideshi Itoh, Patrick Legros, Kieron Meagher, Dilip Mookherjee, Andrew Newman, Markus Poschke, Larry Qiu, Richard Rogerson, John van Reenen, Wing Suen, Chang Sun, and seminar participants at Boston University, Bocconi (EARIE), CUHK, HBS (WMS conference), Hitotsubashi, HKU, HKUST, IUPUI (Midwest International Trade Meeting), NUS, Princeton, Singapore Management University, UPF (Barcelona GSE Summer Forum) for their valuable comments. Financial support from the International Economics Section at Princeton University and the University of Hong Kong is greatly appreciated.

[†]School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong, HKSAR. E-mail: ccfour@hku.hk.

I Introduction

Recent empirical research using firm-level survey data (Bloom and Van Reenen 2007, 2010) has substantiated the existence of large variation in management practices across countries and the importance of management quality for firm performance.¹ In particular, Bloom et al. (2013) documented that improving management quality has positive impact on firm productivity and survival. Among various types of management practices defined by Bloom and Van Reenen (2007), how well a firm monitors its workers is one important category. The research question I ask in this paper is that whether improving management quality (i.e., monitoring efficiency) at the economy level (for all firms) benefits every firm after we take into account general equilibrium effects. Moreover, does this aggregate-level improvement necessarily lead to an improvement in welfare and total output? This paper shows that aggregate improvement in management quality triggers a *selection effect* which favors more hierarchical (i.e., more layers) firms. As a result, less hierarchical firms shrink or exit the market. Moreover, if this selection effect is strong enough, an aggregate improvement in management quality leads to an improvement in welfare and total output.

This paper focuses on a particular type of management technology (i.e., management quality): the ability to monitor and punish misbehaving employees. In what follows, I use MT to denote how effectively the firm monitors and punishes misbehaving workers. I focus on this, as it is an important component of overall management technology and affects firm performance substantially.² In order to focus on monitoring activities of the firm and make the model analytically tractable, I adopt a monitoring-based hierarchy model and incorporate it into an industry equilibrium model with heterogeneous firms (i.e., a closed economy version of Melitz 2003). I treat institutional quality (e.g., labor market conditions and levels of business education) as exogenous from the perspective of a firm, but allow firms to make choices about their internal organization subject to the institution. As a result, how well a firm monitors and punishes misbehaving workers is affected by both the institution and the firm's organizational choices.

This article develops an industry equilibrium model with two sectors. One

¹See Bloom and Van Reenen (2007, 2010), Bloom, Sadun, and Van Reenen (2012), and Bloom et al. (2013) for details.

²Evidence can be found in Bloom and Van Reenen (2007, 2010).

sector is a homogeneous sector which is a perfectly competitive sector with a constant returns to scale technology. I assume that there are no incentive issues inside firms of this sector for simplicity. The other sector which is the main focus of my analysis is a monopolistically competitive sector (the CES sector henceforth). It comprises a continuum of differentiated products with a constant elasticity of substitution (CES) à la Dixit and Stiglitz (1977).³ 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 (i.e., quality) for her product. 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 CES sector earn an expected payoff which equals their exogenous outside option due to free entry.

Firms in the CES sector need to incentivize employees through monitoring, 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 inside the firm (i.e., Calvo and Wellisz 1978, 1979; Qian 1994; Acemoglu and Newman 2002), I assume that the firm sets up a monitoring-based hierarchy consisting of multiple layers. Specifically, the firm allocates workers into different layers to make supervisors monitor their direct subordinates and offer efficiency (i.e., 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 work.

The key result from the firm-level analysis is that 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 and the span of control at each layer.⁴ When the firm wants to produce more, it has to increase the span of control owing to the constraint 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. The firm can add a layer and reduce the span of control to save wage payments to workers at existing layers. However, this comes at the cost of extra wage payments to workers at the new layer. In short, adding a layer

³The purpose of having the homogeneous sector is to endogenize the welfare of workers in a tractable way.

⁴The span of control is defined as the ratio of the number of subordinates to the number of direct supervisors.

is *like* an efficiency-enhancing investment with a fixed cost, which generates *endogenous* increasing returns to scale at the firm level.

I then incorporate this monitoring-based hierarchy model into the closed economy version of Melitz (2003) and evaluate the impact of an economy-wide improvement in the efficiency of monitoring and firing shirking workers. This improvement occurs when labor markets are deregulated, or better management education and practices are introduced into an economy. The key insight is that such an improvement generates a selection effect that favors big firms. An improvement in MT benefits all firms by reducing their labor cost. Importantly, firms with more layers gain *disproportionately* from such a shock, as their average variable costs (AVCs) increase less rapidly with output (i.e., a smaller elasticity of AVC with respect to output). They expand more, since the organizationally augmented production technology features *endogenous* increasing returns to scale at the firm level. This extra expansion creates competitive pressure on the smallest firms and force them to leave the market.

The selection effect of an improvement in MT yields several testable implications. First, the resulting firm size distribution moves to the right in the first-order-stochastic-dominance (FOSD) sense, which is consistent with the finding from Poschke (2014). This result is also consistent with the finding from Hsieh and Klenow (2009, 2012) and Hsieh and Olken (2014) that Mexico and India, whose firms are doing worse at removing misbehaving workers, have more small firms and fewer big firms than the U.S. Finally, surviving firms either *increase* the number of layers or enlarge the span of control, which is consistent with Bloom et al. (2013)'s finding.⁵

Firm's *heterogeneous* choice about the number of layers is the key to generating the selection effect. In a hypothetical world, if all firms had the same number of layers before MT improves *and* cannot change the number of layers after MT improves, this heterogeneous impact would disappear perfectly. In this world, all firms have the same AVC function (both before and after MT improves), and a downward shift of the AVC curve (because of better MT) affects every firm *proportionately*. Note that the selection effect still exists, if firms choose the number of layers optimally before MT improves, but are *not* allowed

⁵Bloom et al. (2013) documented that Indian firms that had received free access to better MT expanded through the establishment of new plants. This finding is consistent with my model's prediction, since establishing a new plant is like increasing the span of control of the firm owner or adding an additional layer between the owner and lower-level production units.

to change the number of layers after MT improves. This is shown in Section IV, and the reason is that the *endogenous* increasing returns to scale technology (induced by the heterogeneous choice about the number of layers) still exists in this case. In short, fixing the number of layers at the *same* value (for all firms) makes the selection effect disappear.

Although I use the monitoring-based hierarchy model to show the selection effect due to tractability, this effect is more general than being a result of a specific hierarchy model. For instance, I conjecture that an improvement in communication technology should generate the same selection effect in the knowledge hierarchy model in industry equilibrium (e.g., Caliendo and Rossi-Hansberg 2012), as firms with more layers use communication technology disproportionately more. The same selection effect induced by better communication technology should also be found in management hierarchy models with delegation in industry equilibrium (e.g., Grobovsek 2014). In short, the key insight that the hierarchical structure of the firm matters for resource allocation and market competition is a new and general lesson we can draw from the research on firm hierarchies.

The aggregate improvement in MT can lead to either an increase or a decrease in welfare and total output, and the key determinant is the strength of the selection effect. After an improvement in MT, firms increase the share of *non-production workers* in employment in order to exploit the improved MT better, as surviving firms add layers of non-production workers (weakly) into the hierarchy. Although this reorganization is optimal for the firm and increases firm profit (coming from lower efficiency wages paid to existing employees), it is not optimal for the economy as non-production workers do not produce output.⁶ In fact, it is a wasteful way to allocate labor resource from the perspective of a social planner. At the same time, thanks to the selection effect, production workers are reallocated to firms with better demand draws after MT improves. As a result, the economy is using *production workers* more efficiently after MT improves, as more production workers are used to produce more valuable goods. Of course, the total effect depends on which force is stronger. When market competition is stiffer (i.e., when goods are more substitutable) in the CES sec-

⁶I assume non-production workers only monitor or supervise subordinates for tractability. However, the same reasoning applies to a more general setting in which non-production workers (e.g., white collar workers) are less involved in producing real goods.

tor, the reallocation effect becomes stronger. Therefore, welfare and total output increase (after MT improves) when market competition is stronger and vice versa.

My work contributes to the literature studying internal firm organization in the industry equilibrium framework. Existing research focuses either on the choice between outsourcing and FDI (e.g., Antràs 2003, 2005; Antràs and Helpman 2004; Alfaro et al. 2016) or on how market price and wealth distribution affect firm boundary and size (Legros and Newman 2008, 2013, 2014; Conconi, Legros, and Newman 2012; Legros, Newman, and Proto 2014). Different from the existing research, this paper focuses on monitoring and incentivizing activities of the firm and shows how the quality of them affects resource allocation. In addition, this paper emphasizes that the potential negative welfare implication of better monitoring technology in efficiency wage models (e.g., Acemoglu and Newman 2002) can be mitigated or even overturned by the selection effect, which is new to the literature.

This paper contributes to the literature on monitoring-based hierarchies (e.g., Calvo and Wellisz 1978, 1979; Qian 1994; Acemoglu and Newman 2002; Meagher 2003) in several ways.⁷ This paper is the first one that incorporates the monitoring-based hierarchy model into the heterogeneous firm framework and solves the market equilibrium. Then, I analyze the heterogeneous impact of improved MT in an industry equilibrium model and uncover a selection effect, which are new to the literature.

This paper is related to the research on knowledge-based hierarchies (e.g., Garicano 2000; Garicano and Rossi-Hansberg 2004, 2006, 2012; Antràs, Garicano, and Rossi-Hansberg 2006; Caliendo and Rossi-Hansberg 2012). A paper that is related to mine is Caliendo and Rossi-Hansberg (2012). The key difference is that the main focus of my paper is to study how effectively firms punish shirking workers affects resource allocation in a closed economy, while Caliendo and Rossi-Hansberg (2012) focused on how the endogenous choice of the number of layers affects gains in aggregate productivity and welfare after trade liberalization. I.e., the selection effect of improved MT is not studied in Caliendo and Rossi-Hansberg (2012). In addition, monitoring-based hier-

⁷Earlier papers on management hierarchies include those by Williamson (1967), Beckmann (1977), and Keren and Levhari (1979) etc. Mookherjee (2013) provides a survey for this literature.

archy model presented in this paper emphasizes the role of frictions (e.g., the moral hazard problem) in determining firm hierarchy and aggregate economic outcomes. I find that welfare and total output can fall after MT improves. In contrast, welfare and aggregate output cannot fall when information and communication technology improves in the knowledge-based hierarchy model in industry equilibrium.⁸

The remainder of the paper is organized as follows. Section two solves the industry equilibrium model. Sections three investigates how an improvement in MT affects firm size and resource allocation within an industry. Section four explains why an explicit modeling of firms' *heterogeneous* choice about their internal organization is essential for the discussion of aggregate impacts of better MT. Section five concludes.

II An Equilibrium Model of Firm Hierarchy Based on Monitoring

In this section, I develop a model 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 subsequently introduce the model into an equilibrium setting and solve the problem of resource allocation in both the product and labor markets.

II.A 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 follow-

⁸In the knowledge hierarchy model, the social planner's solution for resource allocation is the same as the resource allocation in the decentralized economy.

ing Cobb-Douglas utility function:

$$(1) \quad U = \left(\frac{C_c}{\gamma}\right)^\gamma \left(\frac{C_h}{1-\gamma}\right)^{1-\gamma} - I\psi(a_i),$$

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

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

where $y(\theta)$ is the consumption of variety θ , and 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

$$(3) \quad \mathbf{C} \equiv \left(\frac{C_c}{\gamma}\right)^\gamma \left(\frac{C_h}{1-\gamma}\right)^{1-\gamma},$$

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

$$(4) \quad P^\gamma p_h^{1-\gamma} \equiv 1,$$

where

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

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_e 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 normalize the outside option of an entrepreneur to zero. As a result, the expected payoff of entrepreneurs who choose to enter the CES sector equals their entry cost (i.e., f_e) due to free entry of firms in equilibrium.

II.B The Organization of Production

I follow the literature on monitoring-based hierarchies (i.e., Calvo and Wellisz 1978; Qian 1994; Acemoglu and Newman 2002) in modeling the organization of production. 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.

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. 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

$$(6) \quad q = \int_0^{m^r} a(j) dj,$$

where m_T is the measure of production workers and $a(j)$ is the effort level of the j -th unit of labor inputs. Following the convention (i.e., Calvo and Wellisz 1978; Qian 1994), I use a smaller i to denote a higher layer. As a result, the entrepreneur is at layer zero, and 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. Layer T is the lowest layer in the hierarchy and is occupied by production workers.

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

$$(7) \quad w_i - \psi \geq (1 - p_i)w_i,$$

where $p_i(\leq 1)$ is the probability of *catching* and *firing* a shirking worker, and ψ is the disutility of exerting effort. Following Calvo and Wellisz (1978, 1979), I assume that when the worker is caught shirking and gets fired, he loses the job and receives zero wage (and payoff). 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.⁹ First, the bigger the adjusted span of control, the less frequently a subordinate's behavior is checked by his supervisor. This implies a lower probability of *catching* a shirking worker. Second, MT affects the probability of *detecting* workers' misbehavior. Better management rules or help from consulting firms on monitoring activities (e.g., free consulting as in Bloom et al. 2013) lead to easier detection of workers' misbehavior. Finally, 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)).¹⁰ I capture these effects by

⁹The adjusted span of control is the ratio of the number of subordinates to the total input of monitoring effort exerted by the direct supervisors. Formally, it equals $m_i / (\int_0^{m_i-1} a(j) dj)$ where m_i is the number of workers at layer i .

¹⁰Physical technologies such as information and communication technology also affect the

assuming the following functional form for $p(b, x_i)$:

$$(8) \quad p(b, x_i) = \min \left\{ 1, \frac{1}{bx_i(\theta)} \right\},$$

where $x_i(\theta) \equiv m_i(\theta) / \int_0^{m_i-1} a(j) dj$ is the span of control adjusted by supervisors' effort inputs.¹¹ Parameter $1/b$ reflects the *efficiency* of MT. More specifically, the worse the MT is, the bigger the value of b is. In what follows, we assume that MT is not too efficient (i.e., b is not too small) in the sense that

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

In the last section of the paper, I consider an alternative case in which monitoring is costless and $b = 0$.

A firm may want to hire non-production workers, since it wants to economize on the cost to incentivize workers. 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 and justifies the use of production workers only. 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 without the addition of non-production workers. If the firm hires non-production workers who monitor production workers, the incentive-compatible wage paid to production workers will be reduced. However, 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. In total, the trade-off between paying higher wages to existing workers and paying extra wages to workers at new layer(s) determines the optimal choice of the number of layers.

I characterize two optimal choices of the firm before solving the firm's optimal decisions on the other variables, as these two choices are independent of the

probability of *detecting* workers' misbehavior (e.g., better video surveillance and e-mail scanning). This can be captured by the quality of MT, $1/b$, as well.

¹¹As will be shown, 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 discussions are available upon request.

Figure 1: A Firm with Two Layers

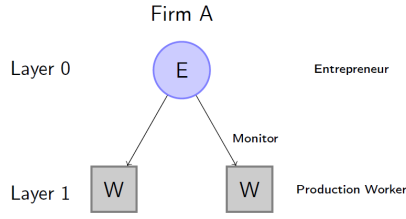
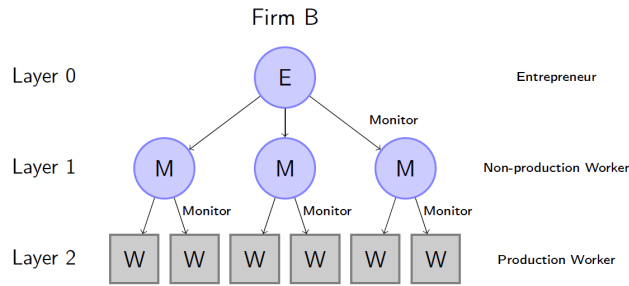


Figure 2: A Firm with Three Layers



firm's decisions on the other variables. Lemma 1 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 A.A. □

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 that *stays* in the market is incentivized to exert effort in equilibrium, since her shirking would cause all her subordinates to shirk, which leads to a negative profit (and payoff) to the firm.

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:

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

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 in the data; see, for example, Rebitzer (1995) and Groshen and Krueger (1990).¹² Based on equations (1), (6), (9) and Lemma 1, the optimization problem for a firm choosing to stay in the market can be stated as

$$(10) \quad \max_{\{m_i\}_{i=1}^T, T} \quad A\theta^{\frac{1}{\sigma}} m_T^{\frac{\sigma-1}{\sigma}} - \sum_{i=1}^T b\psi m_i x_i$$

$$s.t. \quad x_i = \frac{m_i}{m_{i-1}},$$

$$m_0 = 1.$$

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:

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

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 does not exceed one.

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

$$(12) \quad w_T m_T = 2w_{T-1} m_{T-1} = \dots = 2^{T-1} w_1 m_1,$$

where $m_0 = 1$ and $m_T = q$. Thus, the number of workers at layer i is

$$(13) \quad m_i(q, T) = 2^i \left(\frac{q}{2^T} \right)^{\frac{2^T - 2^{T-i}}{2^T - 1}} \quad T \geq i \geq 0.$$

As a result, the firm's span of control at layer i is

$$(14) \quad x_i(q, T) = \frac{m_i(q, T)}{m_{i-1}(q, T)} = 2 \left(\frac{q}{2^T} \right)^{\frac{2^{T-i}}{2^T - 1}} \quad T \geq i \geq 1,$$

which indicates that the number of workers increases disproportionately more

¹²First, Rebitzer (1995) found evidence of a trade-off between supervision intensity and wage payment. Groshen and Krueger (1990) and Ewing and Payne (1999) found evidence on a negative relationship between the span of control and the wage paid to subordinates.

at lower layers when output increases. This is due to the fixed number of entrepreneurs at the top.

Second, optimal output and the number of production workers are determined by

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

where $\beta \equiv (\sigma - 1)/\sigma$. 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

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

and

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

which will be used later. 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.

II.C Layers, Wages, and Empirical Implications

This subsection characterizes the firm's cost functions in order to solve for the optimal number of layers. Furthermore, it derives testable predictions on the relationship between firm size and wages (and the span of control).

Consider a firm that produces q units of output. The variable cost function of such a firm is given by¹³

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

¹³For 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.

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 $T + 1$ layers which can be expressed as

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

Variable cost given the number of layers increases with output. Better MT (a smaller b) pushes down the variable cost given any number of layers proportionately. the following proposition characterizes the AVC curves and the MC curves for the firm.

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 at the output level where the firm adds a layer. The marginal cost falls discontinuously when the firm adds a layer. Firms that produce more have more layers.*

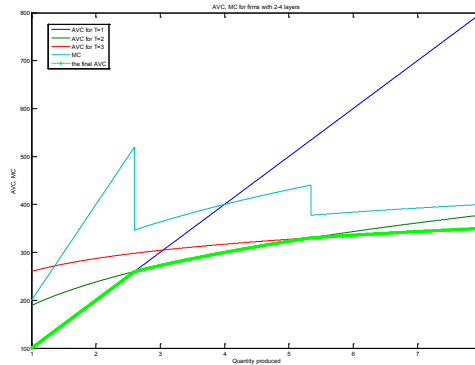
Proof. See Appendix A.B. □

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, and the MC curve is represented by the green curve. The span of control increases at all layers when the firm increases output but not the number of layers, which implies that both AVC and MC increase with output. 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 curve and a kink in the AVC curve when the firm adds a layer.

Proposition 1 establishes a positive relationship between output and the optimal number of layers, which is consistent with the empirical finding of Caliendo, Monte and Rossi-Hansberg (2015). 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 MC at the expense of a fixed cost. In summary, the number of layers and output increase hand in hand in equilibrium.

The relationship between the firm's demand draw and the optimal number of layers is similar to the relationship between output and the optimal number

Figure 3: Average Variable Cost and Marginal Cost



of layers, as the firm with a better demand draw produces more in equilibrium. The following proposition summarizes this relationship.

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

Proof. See Appendix A.C. □

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 the labor cost and incentivizes firms to grow. Moreover, it incentivizes firms with better demand draws to expand more and benefits them more, since the organizationally augmented production technology features *endogenous* increasing returns to scale (with respect to the number of layers). I.e., firms with better demand draws have more layers, and the elasticity of the AVC with respect to output is smaller for these firms. This property is the key to generating the heterogenous effect of improved MT on various firms in an industry equilibrium.

The monitoring-based hierarchy model has predictions for firm-level outcomes, which are summarized by the following proposition.

Proposition 3. *Given the number of layers, both the span of control and wages increase with the firm's quality draw at all layers. Furthermore, the wage ratio (i.e., supervisor's wage divided by direct subordinate's wage) increases with the firm's quality draw at all layers. 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, the wage ratio decreases at existing layers.*

Proof. See Section 2 of Online Appendix. □

The change in the span of control is the key to understanding this proposition. When the firm keeps the number of layers unchanged, the only way to expand is to increase the span of control. When the span of control is larger, higher incentive-compatible wages are needed. Furthermore, wages increase disproportionately more at upper layers, since the share of workers at upper layers (in total employment) decreases when the firm grows without adjusting the number of layers. As a result, the firm tolerates disproportionately more increases in wages at upper layers.

When the firm expands by adding a layer, the constraint at the top is relaxed.¹⁴ 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, as the share of workers at upper (existing) layers (in total employment) increases after the addition of a new layer. This increase incentivizes the firm to reduce wages at upper layers more.¹⁵

The results of Proposition 3 are consistent with the evidence presented in Caliendo, Monte and Rossi-Hansberg (2015). Importantly, the non-monotonic relationship between firm size and wages (and relative wages) finds support from the French data used in Caliendo, Monte and Rossi-Hansberg (2015). Furthermore, the model's unambiguous prediction on the relative wage receives support from findings in Caliendo, Monte and Rossi-Hansberg (2015).

II.D Equilibrium Analysis

I close the model by aggregating across firms and solve for the 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 and not entering in equilibrium due to the large pool of potential entrepreneurs. Workers choose which sector and which labor submar-

¹⁴As 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.

¹⁵Since adding a layer happens discretely when the demand draw increases, it is true in the theory that firms with more layers pay higher wages, if we impose restrictions on functional forms and parameter values.

ket to enter, which will be explained in what follows. They must be indifferent between sectors and various labor submarkets in equilibrium, since they can freely move across sectors and firms.

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 zero payoff can be written as

$$(19) \quad \Pi(\bar{\theta}, A) = 0,$$

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

$$(20) \quad \int_{\bar{\theta}}^{\infty} \Pi(\theta, A)g(\theta)d\theta = f_e,$$

where $g(\theta)$ 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 this sector either. Given the assumptions of a linear production technology and perfect competition, 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

$$(21) \quad p_h L_h = (1 - \gamma)E,$$

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' wage in this sector.

The labor market in the CES sector is characterized by directed 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 (a labor submarket) corresponds to a firm-layer pair (θ, i) , as different firms offer different wages for various positions (i.e., layers). 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. In total, I have

$$(22) \quad \frac{m_i(\theta)}{Q(\theta, i)}(w_i(\theta) - \psi) = \frac{m_{i'}(\theta')}{Q(\theta', i')}(w_{i'}(\theta') - \psi) = p_h \quad \forall (i, i') \forall (\theta, \theta'),$$

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. Different job turn-down rates across labor submarkets (i.e., $\frac{Q(\theta, i) - m_i(\theta)}{Q(\theta, i)} \geq 0$) 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 equation (22), or

$$(23) \quad \begin{aligned} L_c &= \int_{\theta=\bar{\theta}}^{\infty} \sum_{i=1}^{T(\theta, A)} Q(\theta, i) \frac{Mg(\theta)}{1 - G(\bar{\theta})} d\theta \\ &= \frac{WP(\bar{\theta}, A, M) - \psi LD(\bar{\theta}, A, M)}{p_h}, \end{aligned}$$

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 which equals

$$(24) \quad LD = \int_{\theta=\bar{\theta}}^{\infty} \sum_{i=1}^{T(\theta,A)} m(\theta, i) \frac{Mg(\theta)}{1 - G(\bar{\theta})} d\theta.$$

Equation (23) says that *total* expected payoff of workers entering the CES sector (i.e., $p_h L_c$) is the difference between the total wage payment and the total disutility to exert effort.¹⁶

Second, the labor-market-clearing condition indicates that the number of workers employed in the homogeneous sector is

$$(25) \quad L_h = L - L_c.$$

Equations (23) and (25) are two labor market equilibrium conditions that are used to determine the allocation of labor between sectors.

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

$$(26) \quad E = \int_{\bar{\theta}}^{\infty} LC(\theta) \frac{Mg(\theta)}{1 - G(\bar{\theta})} d\theta + p_h L_h + \left[f_0 M + f_e \frac{M}{1 - G(\bar{\theta})} \right] + \psi M,$$

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 (26) is the demand for the final composite good by firms, and the last part of RHS of equation (26) is the consumption of active entrepreneurs who earn profit in equilibrium. Total income of the economy equals total expenditure which includes two parts: demand from workers and demand from firms (i.e., entrepreneurs).¹⁷

The 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 (19), (20), (21), (23), (25) and (26)). One equilibrium condition is redundant due to Walras' law, and I normalize the price of the final composite good to one.

¹⁶For further discussion of the labor market equilibrium in the CES sector, see Section 3 of Online Appendix.

¹⁷Note that only workers and *active* entrepreneurs demand goods in equilibrium. Entrepreneurs who choose not to enter the CES sector and those who enter the CES sector but do not produce do not consume goods.

One implicit assumption for the existence of the equilibrium is that the probability of being employed implied by equation (23) is smaller than or equal to one in *every* labor submarket in equilibrium. In other words, wages offered in the CES sector must satisfy

$$(27) \quad w_i(\theta) - \psi \geq p_h \quad \forall (i, \theta),$$

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. In the paper, I focus on the case in which unemployment exists in *every* labor submarket for two reasons. First, the model yields clean insights and testable implications in this case. Second, firm-level and aggregate-level predictions derived in this case are consistent with existing evidence. The following proposition discusses the existence and uniqueness of an equilibrium with unemployment.

Proposition 4. *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 $\frac{\sigma-1}{\sigma} < \gamma$, there exists a unique equilibrium with unemployment in every labor submarket if the labor endowment is big enough.*

Proof. See Section 3 of Online Appendix. □

The labor endowment affects the outside option of workers through two channels. First, a bigger labor endowment reduces worker's 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 bigger demand for labor and an increase in equilibrium wage in the homogeneous sector. The relative strength of these two effects depends on σ . When $\frac{\sigma-1}{\sigma} > \gamma$ the latter effect dominates, and a small enough labor endowment assures the existence of a unique equilibrium with unemployment in all labor submarkets.

III Management Technology, Institutional Quality, and Aggregate Economic Outcomes

This section investigates how an economy-wide improvement in MT affects firm characteristics, the firm size distribution and welfare.

III.A Selection Effect of Better Management Technology

I consider a scenario in which MT that is common across all firms improves. This improvement occurs when labor markets are deregulated, or better management education and practices are introduced into an economy. 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 cost decreases.

An improvement in MT generates a selection effect that reallocates resources toward bigger and more efficient firms. Specifically, such an improvement benefits all firms, since it reduces firms' labor cost. Moreover, firms with more layers gain more, since the organizationally augmented production technology features endogenous increasing returns to scale (with respect to the number of layers).¹⁸ 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 exit; firms whose demand draws are in the middle receive shrinking revenue and profit; and firms with the best demand draws expand. In short, an improvement in MT facilitates within-industry resource reallocation through benefiting bigger firms more, which is what Bloom et al. (2013) argued in their paper.

Heterogeneity in the number of layers is the key to understanding the selection effect. In a hypothetical world, if all firms were forced to have the same number of layers, this uneven effect would disappear. This is because all firms would operate using the same AVC function in such a world, and the endogenous increasing returns to scale production technology would disappear without the *heterogeneity in the number of layers*. As a result, the exit cutoff and firms' relative market shares would not be affected by improved MT in this hypothetical world. This explains why I adopt a hierarchy model with the endogenous choice of the number of layers in order to study the selection effect of better MT. For more details, see Section IV.

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

¹⁸These firms' AVCs increase less rapidly with output.

with a coefficient k , or

$$(28) \quad G(\theta) = 1 - \left(\frac{\theta_{min}}{\theta}\right)^k,$$

where $G(\theta)$ is the CDF of the demand draw θ . The following proposition summarizes changes in firm characteristics when MT gets improved. 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 5. *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 cut-off 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 A.D. □

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. I focus on this case, as there are always some 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.

The FOSD result for the firm size distribution has implications for resource allocation and are consistent with the data. It implies economies with superior MT are associated with fewer small firms and bigger average firm size. Furthermore, firms with better demand draws have larger market shares in such economies as well. These predictions are consistent with several recent empirical findings. For instance, Hsieh and Klenow (2009, 2012) showed that India and China are worse at getting efficient firms to obtain big market shares compared with the U.S. Hsieh and Olken (2014) found that it is the big firms that are constrained more in developing countries, not the small firms. In short, empirical predictions of the key theoretical result of my model (i.e., the selection effect) square well with the existing evidence.

How does the internal organization of the firm evolve when MT improves? The model predicts that surviving firms either *increase* the number of layers or enlarge the span of control, as better MT incentivizes firms to expand. Bloom et al. (2013) documented that Indian firms that had received free access to better MT expanded through the establishment of new plants rather than the expansion of existing plants. This finding is consistent with my model's prediction, since establishing a new plant is like increasing the span of control of the firm owner or adding an additional layer between the owner and lower-level production units.

III.B Better Management Technology, Welfare and Total Output

Other than firm-level outcomes, I am also interested in how improved MT affects the worker's welfare, which is the worker's expected payoff obtained from entering the CES (or homogeneous) sector. 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. Second, there is monopolistic distortion in one of the two sectors of this economy. Finally, there is a labor market friction due to the directed search. Therefore, a reduction in one friction does not necessarily increase welfare. 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 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 on average. At the same time, firms demand more labor as well.¹⁹ On top of that, the elasticity of substitution governs 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, and this bigger increase eventually raises the aggregate income, making the market size bigger as well. As a result, the CES sector accommodates more firms, and its aggregate labor demand increases, which leads to an increase in the average probability of being employed. In summary, when MT improves and the

¹⁹Remember that average firm size in terms of employment increases when MT improves.

elasticity of substitution is high, the increase in the probability of getting a job dominates the ex post wage loss, which leads to a higher welfare for the worker. The opposite story happens when the elasticity of substitution is low. I.e., the increase in employment per firm is small, which leads to fewer firms and lower aggregate labor demand in the CES sector. Worker's welfare goes down in this case, as both average wage and average employment rate decrease.²⁰

The aggregate improvement in MT can lead to either an increase or a decrease in total output as well, and the key determinant is the strength of the selection effect again. After an improvement in MT, firms increase the share of *non-production workers* in employment in order to exploit the improved MT better, as surviving firms add layers of non-production workers (weakly) into the hierarchy. Although this reorganization is optimal for the firm and increases firm profit (coming from lower efficiency wages paid to existing employees), it is not optimal for the economy as non-production workers do not produce output. In fact, it is a wasteful way to allocate labor resource from the perspective of a social planner. At the same time, thanks to the selection effect, production workers are reallocated to firms with better demand draws. As a result, the economy is using *production workers* more efficiently after MT improves, as more production workers are used to produce more valuable goods. Of course, the total effect depends on which force is stronger. When market competition is stiffer (i.e., when goods are more substitutable) in the CES sector, the reallocation effect becomes stronger, meaning that expansion of firms with better demand draws is large after MT improves. Therefore, total output increases when market competition is stronger and vice versa.

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 (and Table 2) presents an example in which welfare increases (and decreases) when MT improves.

IV Selection Effect and the Number of Layers

In this section, I show that the selection effect disappears, when the number of layers is fixed at a constant for all firms or monitoring becomes costless.

²⁰The 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: Change in Welfare when MT Improves and σ is Big

	Welfare	Ave(wage)	Ave(jtr)	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

Notes: jtr: job turn-down rate; M: the mass of active firms; E: total income
 Parameter Value: $\sigma = 3.8, \gamma = 0.6, \psi = 0.3$

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

	Welfare	Ave(wage)	Ave(jtr)	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

Notes: jtr: job turn-down rate; M: the mass of active firms; E: total income
 Parameter Value: $\sigma = 2.8, \gamma = 0.75, \psi = 0.3$

In Appendix B, I show that the selection effect disappears again, if we do not model the choice of the number of layers. On the contrary, I show that the selection effect still exists, if firms choose the number of layers optimally before MT improves, but *cannot* adjust the number of layers after MT improves. For simplicity, I consider an economy in which the demand draw follows a Pareto distribution as specified in equation (28) in this subsection.

IV.A Fixed Number of Layers (at a Constant for all Firms)

In this subsection, I consider a case in which all firms use the structure of owner-production worker (i.e., $T = 1$) to produce. After substituting the profit function (for $T = 1$) in equation (16) and the ZCP condition in equation (19) into the FE condition in equation (20), I end up with

$$f \int_{\theta=\bar{\theta}}^{\infty} \left(\frac{\theta}{\bar{\theta}}\right)^{\frac{2}{\sigma+1}} \frac{k\theta_{min}^k}{\theta^{k+1}} d\theta - f \left(\frac{\theta_{min}}{\bar{\theta}}\right)^k = f_e,$$

which leads to the result that

$$\bar{\theta} = \theta_{min} \left(\frac{f\left(\frac{2}{\sigma+1}\right)}{f_e \left(k - \frac{2}{\sigma+1}\right)} \right)^{\frac{1}{k}}.$$

Thus, the exit cutoff, $\bar{\theta}$, is *independent* of the quality of MT (i.e., $1/b$). The key to understanding this result is that the ratio of operating profit (and revenue) of

two firms receiving different demand draws is *independent* of b :

$$\frac{\pi(\theta_1, 1)}{\pi(\theta_2, 1)} = \left(\frac{\theta_1}{\theta_2}\right)^{\frac{2}{\sigma+1}}.$$

Furthermore, the above property holds for $T > 1$ as well,²¹ any change in MT affects all firms equally, as long as firms have the same number of layers. Therefore, any economy-wide improvement in MT does not favor one type of firm over another. As a result, the selection effect of improved MT and its impact on the firm size distribution disappear.

IV.B Costless Monitoring

In this subsection, I consider the case in which $b = 0$. In other words, monitoring is costless, and there is no need to set up a monitoring-based hierarchy. The wage determination in this case features a downward rigidity, since the firm has to compensate workers the disutility of exerting effort. In equilibrium, wages offered in the CES sector is the same across firms and equal $\psi + p_h$.²² If the disutility of exerting effort, ψ , is too high, there is unemployment in the CES sector in equilibrium and vice versa. For a firm with the demand draw of θ , its profit is

$$\pi(\theta) = \frac{A^\sigma \theta \beta^{\sigma-1}}{(p_h + \psi)^{\sigma-1} \sigma}.$$

As a result, the FE condition can be written as

$$f \int_{\theta=\bar{\theta}}^{\infty} \left(\frac{\theta}{\bar{\theta}}\right)^k \frac{k \theta_{min}^k}{\theta^{k+1}} d\theta - f \left(\frac{\theta_{min}}{\bar{\theta}}\right)^k = f_e,$$

which leads to the result that

$$(29) \quad \bar{\theta} = \theta_{min} \left(\frac{f}{f_e(k-1)} \right)^{\frac{1}{k}}.$$

²¹For any T , the ratio of the operating profit (and revenue) of two firms with demand draws of θ_1 and θ_2 is $\pi(\theta_1, 1)/\pi(\theta_2, 1) = (\theta_1/\theta_2)^{2^T/(2^T+\sigma-1)}$ which is independent of b .

²²Remember that the net payoff of an employee who works in the CES sector is wage minus the effort cost. Firms in the CES sector have no incentives to offer wages higher than $\psi + p_h$, and workers entering this sector do not accept wages below $\psi + p_h$.

Again, a change in the quality of MT does not affect the exit cutoff and relative market shares of two firms with different demand draws.

The key observation is that when the participation constraint binds which is true when b approaches zero,²³ wage setting is not interrupted by the moral hazard problem of the worker in CES sector. As a result, MT which affects monitoring efficiency does not matter for the determination of wage and firm costs. In other words, when we vary b from an extremely big number to zero, wage setting in the CES sector switches from the one derived from the incentive compatibility (IC) constraint (i.e., equation (9)) to the one derived from the participation constraint (PC) which implies $w = p_h + \psi$. After b enters into the range in which equilibrium wage is pinned down by PC, MT does not play a role in determining aggregate variables. Therefore, MT does not generate the selection effect in the case of costless monitoring.²⁴

IV.C No Adjustment in the Number of Layers after MT Improves

In this subsection, I argue that the selection effect still exists, even if firms cannot adjust their numbers of layers after MT improves. The formal proof for this follows the proof for Proposition 5 closely and is presented in Appendix C. Here, I highlight economic insights coming from the proof. The key to understanding this result is that if firms with worse demand draws weakly benefit from such an improvement, it must be the case that firms with better demand draws *strictly* benefit from such an improvement. If the exit cutoff did not increase after MT improve which implies (weakly) increasing profit for firms with the worst demand draws,²⁵ firms with the best demand draws would strictly benefit.

²³Note that workers in CES sector have to obtain at least $p_h + \psi$ as their expected wage in equilibrium. When b is extremely close to zero, wage derived in equation (9) falls below $p_h + \psi$. As a result, the equilibrium wage, $p_h + \psi$, makes the participation constraint bind and the incentive compatibility constraint slack.

²⁴Firms are heterogeneous and offer different wages in equilibrium. As all workers have the same outside option, it might be the case that wages offered in some firms are based on IC constraint while others are based on PC. Analysis for this type of equilibrium is extremely difficult, as there are two many possible cases to consider. The two types of equilibrium (one in which wage offered by all firms is pinned down by IC constraint and the other one in which wage offered by all firms is pinned down by PC) I consider in the paper yield sharp and economically meaningful insights.

²⁵Note that these firms also have the smallest number of layers in equilibrium.

As a result, the free entry condition would be violated. Therefore, the exit cutoff must go up when MT improves, and firms with the worst demand draws exit the market. On top of that, resources are reallocated to firms with the best demand draws. In other words, the selection effect of an improvement in MT still exists when no firms cannot adjust their numbers of layers after MT improves.

In total, this section shows that it is the heterogeneity in the number of layers (across firms) that matters for the selection effect of better MT. And, the adjustment in the number of layers after MT improves is *not* the key to generating the selection effect.

V Conclusions

This paper uses one canonical approach to modeling the incentive problem inside the firm and incorporates a monitoring-based hierarchy into an industry equilibrium framework to show the selection effect of an improvement in MT. By investigating how MT affects firm characteristics and resource allocation, this paper rationalizes several key empirical findings established in the macro-development literature.

The main contribution of this paper is to explore the selection effect of a common improvement in MT for resource allocation. This effect comes from the heterogeneous impact of such an improvement on firms with various efficiency levels. As a result, resources are reallocated from small firms to big firms, which leads to changes in firm size and welfare. Furthermore, the key to understanding this selection effect is shown to be the endogenous selection of the number of layers. These economic insights and implications open room for future research on management quality and firm organization.

Undoubtedly, much more research remains to be done. First, integrating the knowledge-based hierarchy and the monitoring-based hierarchy into a unified framework is an interesting idea. Second, investigating how MT affects firm boundary (e.g., outsourcing or in-house production) is also an interesting topic to explore. Finally, how import competition and export orientation affect management quality empirically is also worth exploring in the near future.

References:

1. Acemoglu, Daron, and Andrew F. Newman. 2002. "The Labor Market and Corporate Structure." *European Economic Review* 46 (10): 1733-1756.
2. Alfaro, Laura, Paola Conconi, Harald Fadinger, and Andrew F. Newman. 2016. "Do Prices Determine Vertical Integration?" *Review of Economic Studies* 83 (1): 1-34.
3. Antràs, Pol, Luis Garicano, and Esteban Rossi-Hansberg. 2006. "Offshoring in a Knowledge Economy." *Quarterly Journal of Economics* 121 (1): 31-77.
4. Antràs, Pol. 2003. "Firms, Contracts, and Trade Structure." *Quarterly Journal of Economics* 118 (4): 1375-1418.
5. Antràs, Pol. 2005. "Incomplete Contracts and the Product Cycle." *American Economic Review* 95 (4): 1054-1073.
6. Antràs, Pol, and Elhanan Helpman. 2004. "Global Sourcing." *Journal of Political Economy* 112 (3): 552-580.
7. Atkeson, Andrew, and Ariel Burstein. 2010. "Innovation, Firm Dynamics, and International Trade." *Journal of Political Economy* 118 (3): 433-484.
8. Beckmann, Martin J. 1977. "Management Production Function and the Theory of the Firm." *Journal of Economic Theory* 14 (1): 1-18.
9. Bloom, Nicholas, and John Van Reenen. 2007. "Measuring and Explaining Management Practices Across Firms and Countries." *Quarterly Journal of Economics* 122 (4): 1341-1408.
10. Bloom, Nicholas, and John Van Reenen. 2010. "Why do Management Practices Differ across Firms and Countries?" *Journal of Economic Perspectives* 24 (1): 203-224.
11. Bloom, Nicholas, Raffaella Sadun, and John Van Reenen. 2012. "The Organization of Firms across Countries." *Quarterly Journal of Economics* 127 (4): 1663-1705.

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): 1-51.
13. Caliendo, Lorenzo, and Esteban Rossi-Hansberg. 2012. "The Impact of Trade on Organization and Productivity." *Quarterly Journal of Economics* 127 (3): 1393-1467.
14. Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg. 2015. "The Anatomy of French Production Hierarchies." *Journal of Political Economy* 123 (4): 809-852.
15. Calvo, Guillermo, and Stanislaw Wellisz. 1978. "Supervision, Loss of Control, and the Optimum Size of the Firm." *Journal of Political Economy* 86 (5): 943-952.
16. Calvo, Guillermo, and Stanislaw Wellisz. 1979. "Hierarchy, Ability, and Income Distribution." *Journal of Political Economy* 87 (5): 991-1010.
17. Conconi, Paola, Patrick Legros, and Andrew F. Newman. 2012. "Trade Liberalization and Organizational Change." *Journal of International Economics* 86 (2): 197-208.
18. Dixit, Avinash, and Joseph Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67 (3): 297-308.
19. 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 (2): 424-432.
20. Garicano, Luis. 2000. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy* 108 (5): 874-904.
21. Garicano, Luis, and Esteban Rossi-Hansberg. 2004. "Inequality and the Organization of Knowledge." *American Economic Review* 94 (2): 197-202.

22. Garicano, Luis, and Esteban Rossi-Hansberg. 2006. "Organization and Inequality in a Knowledge Economy." *Quarterly Journal of Economics* 121 (4): 1383-1435.
23. Garicano, Luis, and Esteban Rossi-Hansberg. 2012. "Organizing Growth." *Journal of Economic Theory* 147 (2): 623-656.
24. Groshen, Erica, and Alan B. Krueger. 1990. "The Structure of Supervision and Pay in Hospitals." *Industrial and Labor Relations Review* 43 (3): 1348-1468.
25. Grobovsek, Jan. 2014. "Managerial Delegation and Aggregate Productivity." University of Edinburgh, unpublished manuscript.
26. Hsieh, Chiang-Tai, and Pete Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124 (4): 1403-1448.
27. Hsieh, Chiang-Tai, and Pete Klenow. 2012. "The Life Cycle of Plants in India and Mexico." NBER Working Paper 18133.
28. Hsieh, Chang-Tai, and Benjamin Olken. 2014. "The Missing "Missing Middle"." *Journal of Economic Perspectives* 28 (3): 89-108.
29. Keren, Michael, and David Levhari. 1979. "The Optimal Span of Control in a Pure Hierarchy." *Management Science* 25 (11): 1162-1172.
30. Legros, Patrick, and Andrew F. Newman. 2008. "Competing for Ownership." *Journal of the European Economic Association* 6 (6): 1279-1308.
31. Legros, Patrick, and Andrew F. Newman. 2013. "A Price Theory of Vertical and Lateral Integration." *Quarterly Journal of Economics* 128 (2): 725-770.
32. Legros, Patrick, and Andrew F. Newman. 2014. "Contracts, Ownership, and Industrial Organization: Past and Future." *Journal of Law, Economics, and Organization* 30 (suppl 1): i82-i117.
33. Legros, Patrick, Andrew F. Newman, and Eugenio Proto. 2014. "Smithian Growth through Creative Organization." *Review of Economics and Statistics* 96 (5): 796-811.

34. Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71 (6): 1695-1725.
35. Meagher, Kieron J. 2003. "Generalizing Incentives and Loss of Control in an Optimal Hierarchy: the Role of Information Technology." *Economics Letters* 78 (2): 273-280.
36. Melitz, Marc J., and Gianmarco IP Ottaviano. 2008. "Market Size, Trade, and Productivity." *Review of Economic Studies* 75 (1): 295-316.
37. Mookherjee, Dilip. 2013. "Incentives in Hierarchies." In *The Handbook of Organizational Economics*, edited by Robert Gibbons and John Roberts, 764-798. Princeton: Princeton University Press.
38. Poschke, Markus. 2014. "The Firm Size Distribution across Countries and Skill-Biased Change in Entrepreneurial Technology." McGill University, unpublished manuscript.
39. Qian, Yingyi. 1994. "Incentives and Loss of Control in an Optimal Hierarchy." *Review of Economic Studies* 61 (3): 527-544.
40. 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 (1): 107-129.
41. Shapiro, Carl, and Joseph Stiglitz. 1984. "Equilibrium Unemployment as a Worker Discipline Device." *American Economic Review* 74 (3): 433-444.
42. Williamson, Oliver. 1967. "Hierarchical Control and Optimum Firm Size." *Journal of Political Economy* 75 (2): 123-138.

A Appendix: Proofs

A.A 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 to before, removing them from the hierarchy does not affect the firm's output and (weakly) reduces the labor cost, which means excluding them from the hierarchy is always optimal. Thus, all workers are incentivized to work in equilibrium. Second, the reason why the firm wants to allocate the monitoring 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

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

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

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

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. \square

A.B Proof of Proposition 1

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

$$(A.1) \quad AVC(q, T) = \left(2 - \frac{1}{2^{T-1}}\right)b\psi 2^{1-\frac{T}{2^{T-1}}} q^{\frac{1}{2^{T-1}}}$$

and

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

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 2. 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

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

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

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

Thus, the sign of the above expression depends on

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

or

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

or

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

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

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

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 $(1 - 2^{-(2T+1)}) \ln\left(\frac{2^{T+1} - 1}{2^T - 1}\right) - \ln 2$ with respect to T which has just been proved, I conclude that

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

and

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

Therefore, q_T must be an increasing function of T . \square

Having established the monotonicity of q_T in Lemma 2, I can characterize the overall shape of the AVC curve.

Lemma 3. *If the output produced in equilibrium $q \in [q_{T-1}, q_T)$, the optimal number of 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 proves 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. \square

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 3. 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. \square

In sum, I proved Proposition 1 due to Lemma 3 and Claim 1. \square

A.C Proof of Proposition 2

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

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

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

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

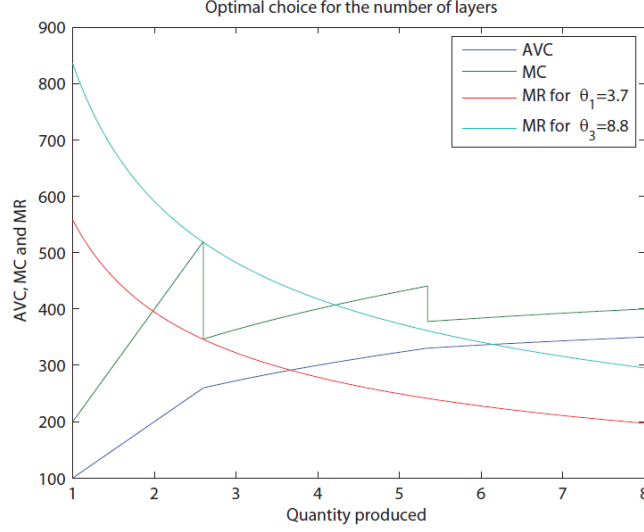
$$MR(\theta_{T3}, q_T) = A\beta\theta_{T3}^{\frac{1}{\sigma}} q_T^{-\frac{1}{\sigma}} = MC(q_T, T) = b\psi 2^{2 - \frac{T}{2^T-1}} q_T^{\frac{1}{2^T-1}}.$$

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

The graphical representation of θ_{T1} and θ_{T3} is in Figure 4.

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.

Figure 4: Lower and Upper Bounds on Layer-Switching from $T = 1$ to $T = 2$



Lemma 4. 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

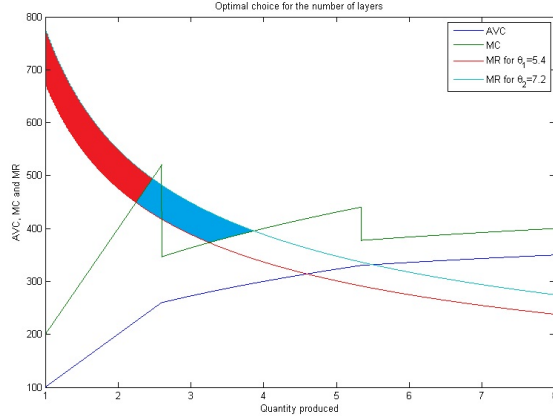
$$\pi(\theta_{T1}, T) \equiv \pi(\theta_{T1}, T, q(\theta_{T1}, T)) > \pi(\theta_{T1}, T, q_T) = \pi(\theta_{T1}, T+1, q_T) = \pi(\theta_{T1}, T+1),$$

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

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

□

Figure 5: Optimal Choice of the Number of Layers



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

Lemma 5. 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 5 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 $q \geq q_T$, 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 $\theta \in [\theta_{T_1}, \theta_{T_3}]$. The continuity of $\pi(\theta, T + 1) - \pi(\theta, T)$ in θ is straightforward to see. \square

Fourth, I prove the following result which is the key step to prove this proposition. More specifically, there exists a threshold $\theta_{T_2} \in (\theta_{T_1}, \theta_{T_3})$ such that firms with this quality draw is indifferent between having $T + 1$ layers and having $T + 2$ layers. Claim 2 summarizes the results.

Claim 2. *For each T , there exists a threshold $\theta_{T_2} \in (\theta_{T_1}, \theta_{T_3})$ such that firms with this demand draw is indifferent between having $T + 1$ layers and having $T + 2$ layers. Moreover, firms with demand draws smaller than θ_{T_2} strictly prefer $T + 1$ layers over $T + 2$ layers, while firms with demand draws bigger than θ_{T_2} strictly prefer $T + 2$ layers over $T + 1$ layers.*

Proof. From Lemma 4, I have

$$\pi(\theta_{T_1}, T) > \pi(\theta_{T_1}, T + 1),$$

and

$$\pi(\theta_{T_3}, T) < \pi(\theta_{T_3}, T + 1).$$

As $\pi(\theta, T + 1) - \pi(\theta, T)$ continuously increases in θ for $\theta \in [\theta_{T_1}, \theta_{T_3}]$ due to Lemma 5, there must exist a threshold $\theta_{T_2} \in (\theta_{T_1}, \theta_{T_3})$ such that

$$\pi(\theta_{T_2}, T) = \pi(\theta_{T_2}, T + 1).$$

And for all $\theta < \theta_{T_2}$

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

while for all $\theta > \theta_{T_2}$

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

\square

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

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

$$MR(\theta_{T_0, T_1}, q_{T_0, T_1}) = MC(q_{T_0, T_1}, T_1),$$

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

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

Second, using the same approach used in the proof of Claim 2, one can prove that there exists a quality cutoff $\theta_{2_{T_0, T_1}} \in (\theta_{T_0, T_1}, \theta_{1_{T_0, T_1}})$ such that firms with quality draws bigger than $\theta_{2_{T_0, T_1}}$ prefer $T_1 + 1$ layers over $T_0 + 1$ layers and 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. From 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.

Thanks to this proposition, I only need to derive the sequence of θ_{T_2} for $T = 1, 2, 3, \dots$ 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. \square

A.D Proof of Proposition 5

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

I make the following notations. Suppose b decreases from 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 6. *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 $b_2 < b_1$. However, a decreasing exit cutoff plus a weakly increasing adjusted market size violate the FE condition defined in Equation (20). Thus, it must be true that $A_2 < A_1$. Second, if $A_2/A_1 \leq b_2/b_1$, 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

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

□

Second, I show that all firms increase the number of layers weakly. It is straightforward to observe that if $A_2/A_1 = b_2/b_1$, the optimal output, employment, and the number of layers would be unchanged. As I have proved that $A_2/A_1 > b_2/b_1$ in Lemma 6, 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 $\bar{\theta} \equiv \bar{\theta}_2/\bar{\theta}_1$, $A \equiv A_2/A_1 < 1$, and $b \equiv b_1/b_2 > 1$. As $\bar{\theta}_2 \leq \bar{\theta}_1$,

$$Y(A, T_0)Z(b, T_0) \geq 1,$$

which implies that

$$(A.3) \quad \frac{b^{(\sigma-1)(2^{T_0}-1)}}{\left(\frac{1}{A}\right)^{2^{T_0}\sigma}} > 1.$$

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. To see this, first note that

$$\frac{\pi(\theta, T(\theta, A_2, b_2), A_2, b_2)}{\pi(\theta, T(\theta, A_1, b_1), A_1, b_1)} = \left(\frac{b^{(\sigma-1)(2^{T(\theta)}-1)}}{\left(\frac{1}{A}\right)^{2^{T(\theta)}\sigma}} \right)^{\frac{1}{\sigma+(2^{T(\theta)}-1)}},$$

where $T(\theta) \equiv T(\theta, A_1, b_1) = T(\theta, A_2, b_2)$ and $T(\theta) \geq T_0$ as $\theta \geq \bar{\theta}_1$. Next, we know both b and $1/A$ are bigger than one. Thus, equation (A.3) implies that

$$\begin{aligned} & \frac{b^{(\sigma-1)(2^{T(\theta)}-1)}}{\left(\frac{1}{A}\right)^{2^{T(\theta)}\sigma}} \\ &= \frac{\left(b^{(\sigma-1)(2^{T_0}-1)}\right)^{\frac{2^{T(\theta)}-1}{2^{T_0}-1}}}{\left(\left(\frac{1}{A}\right)^{2^{T_0}\sigma}\right)^{\frac{2^{T(\theta)}}{2^{T_0}}}} \\ &> 1. \end{aligned}$$

Accordingly, we have

$$\left(\frac{b^{(\sigma-1)(2^{T(\theta)}-1)}}{\left(\frac{1}{A}\right)^{2^{T(\theta)}\sigma}} \right)^{\frac{1}{\sigma+(2^{T(\theta)}-1)}} > 1.$$

If the firm endogenously changes the number of layers, its profit must be bigger than the profit it earns when $b = b_1$ due to the revealed preference argument. In total, I have

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

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

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

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 $1/b$. Based on the above notations and the Pareto distribution on θ , I have

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

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

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

where $T \geq T_0$. 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

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

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

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

where $A \equiv A_2/A_1 < 1$, and $b \equiv b_1/b_2 > 1$. This expression can be simplified

further to

$$(A.4) \quad \frac{\theta_{T,2}|_{b=b_1}}{\theta_{T,2}|_{b=b_2}} = (bA)^\sigma.$$

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

$$(A.5) \quad \frac{\bar{\theta}_1}{\bar{\theta}_2} = A^\sigma b^{(\sigma-1)(1-\frac{1}{2T_0})}.$$

Finally, from equations (A.4) and (A.5), I conclude that

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

Therefore, for all $T \geq T_0$, $Prob(t > T, b_2) > Prob(t > T, b_1)$ 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 $S(\bar{\theta}_2, A_2) = S(\bar{\theta}_1, A_1) = f / \left[1 - \frac{\beta(2^{T_0}-1)}{2^{T_0}} \right]$ and

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

Based on Equation (A.5), I conclude that

$$(A.6) \quad \frac{S(t\bar{\theta}_2, A_2)}{S(t\bar{\theta}_1, A_1)} = \left[\left(\frac{1}{b} \right)^{(\sigma-1)(1-\frac{1}{2^{T_0}})} b^\sigma \right]^{(\sigma-1)\frac{2^{T_1}-2^{T_0}}{(\sigma+2^{T_1}-1)(\sigma+2^{T_0}-1)}} > 1.$$

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 $T_1 = T_0$ due to Equation (A.6). 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 proved 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 $T(\bar{\theta}_1, A_1) = T(\bar{\theta}_2, A_2) = T_0$,

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

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

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

and

$$b_1 > b_2.$$

Second, Based on the above result I derive that

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

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

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

and

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

where $V2(t, T_0, T_1)$ is a function of (t, T_0, T_1) . Based on equation (A.5), I conclude that

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

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

$$(A.7) \quad SC_i(T, q(\theta, b, A, T(\theta, b, A))) = \frac{m_{i+1}(T, q(\theta, b, A, T(\theta, b, A)))}{m_i(T, q(\theta, b, A, T(\theta, b, A)))},$$

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 as²⁶

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

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.

B Appendix: Selection Effect and the Number of Layers

In this section, I show that the selection effect disappears, if we ignore the organizational choice of the firm and interpret the economy-wide improvement in MT as a proportional upward shift of firms' demand draws in Melitz (2003). Specifically, suppose every demand draw increases from θ to $(b_1\theta)/b_2$ where $b_1 > b_2$. This is equivalent to moving the distribution of θ from

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

to

$$G(\theta) = 1 - \left(\frac{\frac{b_1}{b_2}\theta_{min}}{\theta} \right)^k,$$

where productivity-enhancing MT improves from $1/b_1$ to $1/b_2$. Equation (29) shows that the exit cutoff increases by the same proportion (i.e., b_1/b_2), and rel-

²⁶For detailed proof of this result, see Appendix A.D.

ative profit (and sales) of two firms with different demand draws is unchanged:

$$\frac{\pi(\theta_1)}{\pi(\theta_2)} = \frac{\pi\left(\frac{b_1}{b_2}\theta_1\right)}{\pi\left(\frac{b_1}{b_2}\theta_2\right)} = \frac{\theta_1}{\theta_2}.$$

Therefore, modeling an improvement in MT in such a reduced-form way does not generate the selection effect.

An alternative benchmark model for heterogeneous firms is Melitz and Ottaviano (2008). We show that the selection effect of improved MT is reversed (i.e., favoring small and less hierarchical firms), if we interpret the economy-wide improvement in MT as a proportional downward shift of firms' cost draws in Melitz and Ottaviano (2008). Suppose the cost draw of every firm changes from c to $(b_2c)/b_1$ where $b_1 > b_2$, when MT improves. As a result, the CDF of the cost draw becomes²⁷

$$G(c) = \left(\frac{c}{\frac{b_2}{b_1}c_M}\right)^k.$$

After this change, the exit cutoff changes from $c_D(b_1)$ to $c_D(b_2) = (b_2/b_1)^{k/(k+2)}c_D(b_1) > (b_2/b_1)c_D(b_1)$, where $c_D(b_i)$ is the exit cutoff when the management quality is $1/b_i$ (see equation (15) of Melitz and Ottaviano 2008). Therefore, the exit cutoff on the cost draw goes down by less than b_2/b_1 (i.e., less stringent selection after MT improves), and relative profit of two firms with different cost draws ($c_1 > c_2$) equals

$$\frac{\pi(c_1)}{\pi(c_2)} = \frac{(c_D(b_1) - c_1)^2}{(c_D(b_1) - c_2)^2}$$

before the improvement of the MT and

$$\frac{\pi\left(\frac{b_2}{b_1}c_1\right)}{\pi\left(\frac{b_2}{b_1}c_2\right)} = \frac{\left(c_D(b_2) - \frac{b_2}{b_1}c_1\right)^2}{\left(c_D(b_2) - \frac{b_2}{b_1}c_2\right)^2}$$

after the improvement. As $c_D(b_2) > (b_2/b_1)c_D(b_1)$, less efficient firms gain more from such a proportional downward shift in the cost draw:

$$\frac{\pi\left(\frac{b_2}{b_1}c_1\right)}{\pi\left(\frac{b_2}{b_1}c_2\right)} > \frac{\pi(c_1)}{\pi(c_2)}.$$

²⁷See equation (14) of Melitz and Ottaviano (2008).

In other words, smaller and less hierarchical firms survive and gain market shares (relative to more hierarchical firms) after MT improves, which leads to a selection effect favoring *small* firms. This result is related to the difference in markups between firms. A downward shift in the cost draw increases profit of all firms. Importantly, as more efficient firms charge higher markups, this proportional downward shift in the cost draw increases the profit of less efficient firms disproportionately. Therefore, the selection effect favoring more hierarchical firms is reversed in this case. In total, the selection effect favoring big firms cannot be generated, if we do not model the firm's hierarchical choices explicitly.

C Appendix: No Adjustment in the Number of Layers

In this section, we show that the selection effect (i.e., the increase in the exit cutoff after MT improves) still exists, if firms can choose the number of layers optimally before MT improves but *cannot* change it after MT improves. In short, the proof for this result resembles Appendix A.D, and we prove it using the following three steps. First, using the same logic as in Appendix A.D, we can show that

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

where $1/b_1$ and A_1 are management quality and adjusted market size before MT improves, while $1/b_2$ and A_2 are management quality and adjusted market size after MT improves. Next, to the contrary, we assume that there is no selection effect when there is no adjustment in the number of layers after MT improves. This means that the exit cutoff $\bar{\theta}$ either decreases or stays unchanged when MT improves which implies that

$$(C.1) \quad \pi(\bar{\theta}, T0, A_2, b_2) \geq \pi(\bar{\theta}, T0, A_1, b_1),$$

where $T0 + 1$ is the number of layers chosen by firms on the exit cutoff before MT improves, and subscript 1 (and 2) denotes the variable before (and after) MT improves. Following the same steps used in Appendix A.D, we can show

that

$$\frac{b^{(\sigma-1)(2^{T_0}-1)}}{\left(\frac{1}{A}\right)^{2^{T_0}\sigma}} \geq 1,$$

where $A(< 1)$ and $b(> 1)$ are defined in the same way as in appendix A.D. Furthermore, the above inequality implies that for all other firms whose number of layers $(T + 1)$ is bigger than or equal to $T_0 + 1$ we also have

$$\frac{b^{(\sigma-1)(2^T-1)}}{\left(\frac{1}{A}\right)^{2^T\sigma}} > 1,$$

which implies that

$$(C.2) \quad \pi(\theta, T, A_2, b_2) > \pi(\theta, T, A_1, b_1),$$

where $T \geq T_0$ and $\theta \geq \bar{\theta}$. Equations (C.1), (C.2) and the assumption that the exit cutoff does not increase (after MT improves) together imply that the free entry condition is violated. Therefore, the exit cutoff $\bar{\theta}$ must go up when MT improves for all firms. As a result, firms with the worst demand draws exit the market, and resources are reallocated to firms with the best demand draws. \square