Specialization, Agency Cost and Firm Size

Sungbin Cho*†

Korea Development Institute

Abstract

This paper extends the principal-agent model to determine the size of the firm as measured by the number of agent hired. Hiring more agents results in benefits and costs to the principal. The benefits are gains from specialization: higher productivity can be achieved if, as the number of agents increases, their task assignments become more specialized. However, increases in task specialization make monitoring more difficult and costly. In this paper I study peer monitoring among agents. Balancing productivity gains with monitoring costs determines the optimal size of the firm. This paper shows that agency costs due to moral hazard are one factor that sets limits on firm size in a model where it would otherwise be unbounded.

JEL codes: D23, L23, M54

*Contact: Korea Development Institute, P.O. Box 113, Cheongyang, Seoul 10-012, Korea. E-mail: scho@kdi.re.kr

†I would like to thank Joseph M. Ostroy for his unceasing guidance and encouragement. I am also grateful to Alberto Bennardo, Hongbin Cai, David K. Levine, Ichiro Obara, Jean-Laurent Rosenthal, and Leeat Yariv for their helpful comments and suggestions. Thanks also go to seminar participants at various seminars. All remaining errors are mine.

1 Introduction

The division of labor and specialization have long been recognized as sources of productivity improvements and increase in economic welfare. In a world where markets could successfully coordinate different specialized tasks, each production unit would specialize in a single stage of production and all possible gains would be realized among highly specialized firms. However, as Coase (1937) notes, transaction costs for using the market are not nil and a firm typically performs more than one task in order to save on transaction costs.[1] Intrafirm gains from specialization can be achieved when the firm performs more than one task and hires more than one agent. A natural question to ask is why a firm does not grow ever larger if gains from specialization can be achieved within the firm. This paper uses the principal-agent framework to describe how transaction costs within a firm bound its size.

Specifically, this paper addresses the question of "Why is the size of the firm, measured by the number of workers, bounded when increasing returns from specialization prevail?" This paper focuses on agency costs, rather than aspects of technology, as a factor that sets limits on firm size.

Capitalizing on recent developments in the repeated agency literature, this paper proposes a model that explicitly takes into account the trade-off between gains from specialization and agency costs within a firm. When hiring more than one agent, there are costs and benefits. The benefits are captured through gains from specialization. Among the varieties of gains from specialization, this paper tries to capture two types in particular. The first one is economies of scale: the divisions of a complex job into many simpler tasks enable workers to be more skilled and hence productive to a degree that otherwise would be impossible. The second type of gain stems from the fact that the division of labor makes it possible to save on fixed costs that are commonly incurred when agents pass from one type of work to another. Therefore, if agents' task assignments become more specialized as more agents are hired, then higher productivity gains and saving on fixed costs can be achieved.

The costs of hiring many agents are manifested by the monitoring costs. In this paper, I think of monitoring costs as a function not only of the number of agents but also of task assignments. When the firm has workers perform the same tasks, I expect the cost of monitoring not to rise significantly as more agents are hired. However, in this case there is no gain from specialization. On the other hand, when the firm assigns different tasks to

[1] For example, Enright (1995) reports in his case study of the Swiss watch industry that high transaction costs give rise to an integration of tasks and specialization occurs within the firm.

different agents, the gains from specialization increase but at the same time the difficulty and cost of monitoring also rises. Under some mild assumptions, this paper shows that monitoring costs can dominate gains from specialization. As a result, the size of the firm is bounded.

This paper considers 'peer' (or mutual) monitoring instead of the vertical control system used in the standard principal-agent model. Rather than the principal or supervisor doing the monitoring, the agents monitor one another. Principal monitoring and supervision is appropriate in cases where joint production is not important and each agent's contribution can be easily identified. But when the principal can only observe the final outcome of joint production and when it is too costly to distinguish individual performance through vertical control, the principal is better off relying on peer monitoring.[2]

Peer monitoring has long been recognized as a way to control agency problems. Though not formally developed, the role of peer monitoring is emphasized by Alchian and Demsetz (1972), Williamson (1975), Fama and Jensen (1983), Nalbantian (1988), Levine and Tyson (1990) and Kandel and Lazear (1992). Fama and Jensen (1983) claim that "when agents interact to produce output, they acquire low-cost information about colleagues, information not available to higher-level agents. Mutual monitoring systems tap this information for use in the control process."

Peer monitoring is not just a modeling device. It is often observed in reality. One might think that peer monitoring is appropriate for partnership but not for large corporations. However, mutual monitoring often occurs in large corporations. There is much anecdotal evidence on how Japanese firms use the mutual monitoring schemes. In the US, Knez and Simester (2000) examine the case of Continental Airlines, a firm with about 35,000 employees. They claim that mutual monitoring, when combined with a bonus scheme, has enhanced productivity. Therefore, though peer monitoring is not widely used in agency models, it is not unrealistic.[3]

[2]The literature on principal/multi-agent model (e.g. Ma (1988)) and on group-lending model (e.g. Arnott and Stiglitz (1991)) assumes that only peer monitoring generates informative signals.

[3]Though I do not consider hierarchies, peer monitoring is not incompatible with hierarchies. That is, in some cases peer monitoring is an abstraction of the hierarchical structure. For example, consider a firm consisting of several divisions. Each division has a division head (a supervisor) who reports to the principal. The report from each division, which is costly to prepare, contains information about other divisions. For example, a report from the marketing division may contain information about product quality, cost of production, etc. Based on the report from the marketing division, the principal can know or infer performances of the other divisions. Ignoring the possibility of collusion among the divisions, and assuming the report from each division is informative about performance, it is possible to compress the hierarchical

The tool I use to explore the issue of firm size is a repeated principal/multi-agent model. I adopt the multi-agent approach to capture the gains from specialization. Moreover, I think of a firm as a group of people working together over time, which leads to use the repeated agency framework. When all public information is contractible, I may not need to consider repeated agency. However, where some publicly observable variables are not contractible, repeated agency is an appropriate framework because the principal relies on implicit contracts to provide incentives.

The main interest of this paper lies in the trade-off between gains from specialization and monitoring costs. Therefore, this paper sets aside the issue of asset ownership.[4] This paper also sets aside the issues of how firms form and who becomes a principal. It is assumed that a firm exists and the set of tasks that the firm performs is given. But, ignoring asset ownership does not mean that the ownership structure is not important in determining the size of the firm. In some environments, asset ownership is crucial for providing incentives in the presence of incomplete contracts and may be a driving force in determining the size of the firm, as Grossman and Hart (1986) and Hart and Moore (1990) (GHM) note.[5]

The reason for not considering asset ownership is that this paper focuses on the entrepreneurial role of the principal. That is, the primary role of the principal is organizing production. This view goes back to Clark (1909) who notes that "the entrepreneur function in itself includes no working and no owning of capital: it consists entirely in the establishing and maintaining of efficient relations between the agents of production." Kaldor (1934) echoes Clark and suggests that the fixed factor limiting the size of the firm is coordination. In my view, the rents the principal receives come from the ownership of intellectual capital for organizing production, not from physical capital.[6] In this sense, the principal is an

structure of the firm into a flat one with peer monitoring.

[4]One justification comes from Hansmann (1996) who notes that

> a business corporation is just a particular type of cooperative: a cooperative is a firm in which ownership is assigned to a group of the firm's patrons, and the persons who lend capital to a firm are just one among various classes of patrons with whom the firms deal. Conversely, supplying capital to the firm is simply one of many transactional relationships to which ownership can be tied, and there is nothing very special about it. Ownership of a firm need not, and frequently does not, attach to investment of capital. *Indeed, contrary to some popular perceptions and even to some more sophisticated organizational theory, ownership of the firm need have nothing to do with ownership of capital, whether physical or financial.* - Emphasis added

[5]Extending the GHM model, Rajan and Zingales (2001) show that if there are severe expropriation problems by the agents, then residual control rights are important in providing incentives by controlling access to critical resources and, thus, determining the size of the firm.

[6]In this respect, my model is similar in the spirit to Rajan and Zingales (1998, 2001) who consider human

entrepreneur who can be an employee.[7]

The paper is organized as follows. In section 2, I briefly review some of the related literature. Then I present the basic model in section 3, followed by preliminary results on the multilateral relational contract. I introduce the detailed model incorporating gains from specialization and present the main results in the following sections. I consider a binary choice of monitoring in section 4. I assume that a (perfect) signal of production effort is available only when the required level of monitoring effort is made. Also I consider the case where monitoring effort is observable to the principal. Admittedly the assumption of perfect monitoring with observable monitoring effort is rather strong. But this assumption helps reveal the basic tradeoff of hiring more agents. Moreover, assuming perfect monitoring with observable monitoring effort seems a reasonable first step toward analyzing a more general environment of imperfect monitoring and/or unobservable monitoring effort. In section 5, I extend my analysis to the case where the monitoring choice is not binary. Hence, an imperfect signal on production effort is generated through peer monitoring. Depending on the observability of monitoring effort, I consider two cases: observable and unobservable monitoring. In section 6, I make some remarks on extensions and conclude. Proofs of the results are contained in the appendix.

2 Related Literature

Knight (1921) raised the issue of firm size and mentioned the relation between efficiency and size. He stated that "continuous and unlimited expansion of the firm must be offset by some equally powerful one making for decreased efficiency with growth in size."[8] Then Coase (1937) argued that optimal firm size is determined by a comparative assessment of the costs of internalizing additional transactions and the costs of market transactions. Since Coase, however, the literature to date, with a few exceptions, has not dealt formally with the important question of firm size or the number of agents the principal should employ.

To the best of my knowledge, the first paper formally addressing this issue is Williamson (1967). Williamson considers a hierarchical organization within which the workers on the

capital (that might be alienable) as a source of rent to the principal.

[7]Without owning physical assets, the employed principal may have conflicts of interest with the owner of the firm. I do not consider this, not because it is unimportant, but because it complicates matters without adding much to the analysis.

[8]Recited from Coase (1937).

lowest level of the hierarchy carry out all production. He assumes a control loss when moving from the supervisor to the subordinate in which only α $(0 < \alpha < 1)$ fraction of the work done by a subordinate contributes to the objectives of his superior. Williamson's paper implicitly assumes increasing average cost due to the supposition that managers are not directly productive. Moreover, Williamson's main concern is the optimal number of levels in the firm hierarchy rather than what limits the size of the firm. His span-of-control approach asks how to optimize the firm's structure given (and assumed) organizational diseconomies of scale.

Calvo and Wellisz (1978) extend Williamson's analysis and show that if the agents do not know when they are monitored and if it is profitable for a hierarchical firm to arise at all, then control loss cannot limit the size of the firm. A limit on the firm size can exist only if the agents know when they will be monitored. Their result, however, hinges on a critical assumption that as the number of agents increases, the moral hazard problems do not get worsened so much that firm size is bounded.[9]

Aron (1988) and Ziv (1993) examine firm size in the presence of moral hazard. Aron analyzes a model of firm diversification and shows that agency costs coupled with a span-of-control managerial technology determine firm size. But Aron's approach is about the scope of the firm, rather than the size of the firm in a sense that her model is about the question of how many product lines the firm has. In addition, she does not pay attention to the gains from specialization since her model is not a multi-agent one. Ziv's paper is the closest to mine in the sense that he uses a multi-agent model. He uses a linear compensation model developed by Holmström and Milgrom (1987) and compares the firm size in the first-best with that in the second-best. In Ziv's model, however, the first-best firm size is bounded due to the absence of gains from specialization.

Relatedly but independently, several papers have developed arguments to explain why there exist organizational diseconomies of scale. Rasmusen and Zenger (1990) use a statistical inference model to find that small teams can write more efficient incentive contracts than large teams when agents choose individual effort levels but the principal observes only the joint output. They show that the firm makes more errors (the sum of type I and II errors) as the number of employees increases. Their analysis suggests that reductions in the errors can be achieved by monitoring but they do not raise this possibility.

[9]Camacho and White (1981) point this out and show that in the Calvo-Wellisz model, the firm may not be ever formed under this assumption.

McAfee and MacMillan (1995) examine organizational diseconomies of scale by looking at cost of hierarchy in terms of a hidden information model.[10] They show that due to private information, (virtual) marginal cost in the hidden information model is higher the taller the hierarchy. The logic behind their results is similar to the successive marginalization in vertical relationships. Though they recognize that increasing informational rents limit the size of the firm, they do not consider the possibility of monitoring as a way of controlling informational rents.

The above papers recognize the importance of the management cost in determining the size of the firm. But, they do not recognize the importance of intrafirm division of labor and increasing returns as an offsetting factor.

There are several papers that deal with the division of labor. Rosen (1978, 1983) focuses on production substitutability and human capital accumulation in relation to the division of labor. Becker and Murphy (1992) consider the division of labor and coordination cost. Though they notice the importance of coordination costs, free-rider problem and agency costs, they do not consider the moral hazard problem explicitly. Moreover, Becker and Murphy's interest lies in whether the degree of specialization covaries with the extent of the market. Hence their focus is different from mine. Matsui and Postlewaite (1999) examine the job assignment of heterogeneous agents in the presence of gains from specialization. The focus of their paper, however, is on the distribution of income. Overall, these papers do not consider the agency problem with the division of labor.

The issue of firm size is related to the boundaries of firm. As Holmström and Roberts (1998) point out, however, the literature on the boundary issue focuses on the hold-up problem. The main interest of this literature lies in the contractual solutions to asset specificity, opportunism, and merger-or-not decisions. This literature focuses on the make-or-buy decision and implicitly analyzes how many tasks should be carried out in one firm. Little attention has been paid to the issue of firm size *per se*.

Apart from firm size issue, this paper is related to two other strands of literature. One is the multi-agent problem and the other is repeated agency with implicit contracts. Since Holmström (1982), many papers have examined team production and the principal-multiagent models to resolve the moral hazard problem. Specifically, it is known that when agents can engage in costless and perfect monitoring, then certain mechanisms may achieve the first-best (see Ma (1988) and Mookherjee (1984)) in a single period principal-multiagent

[10]Keren and Levhari (1983) also develop similar points.

relationship. The results of these models hinge on two assumptions: the possibility that the principal can use multi-stage mechanisms and costless and perfect monitoring. Moreover, the problem of multiple equilibria arises in such mechanisms. This paper assumes away that the principal can use multi-stage mechanisms as in Ma but considers costly monitoring. In addition, this paper considers imperfect monitoring as well as perfect monitoring.

With regard to repeated agency with implicit contracting, this paper is related to Bull (1987), MacLeod and Malcomson (1988,1989,2000), Pearce and Stacchetti (1998), Che and Yoo (2001), Levin (2002a, b), and Rayo (2002). These papers show the existence of self-enforcing contracts and characterize the set of contracts and equilibrium payoffs in the repeated agency model. Among them, only Che and Yoo (2001), Levin (2002a), and Rayo (2002) consider the repeated multi-agent model in which the number of agents is exogenously given. I depart from these papers by allowing the principal to decide on the firm's optimal size in a moral hazard setting.

The structure of the model is close to Levin (2002a, 2002b) and Rayo (2002). In fact, the results on multilateral relational contracts are extensions of their findings. The novel feature of this paper is that this paper explicitly considers the trade-off between gains from specialization and monitoring costs.

3 Preliminaries

In this section, I describe the basic model and the time line of the game. Then I present preliminary results on multilateral relational contracts. The purpose of this section is to provide a rationale for confining my attention to a stationary contract. I show the optimality of a stationary relational contract, 'given' the number of agents. The results are extensions of the works by Levin (2002a, 2002b) and Rayo (2002).

3.1 Basic Model

In this paper, I consider the employment relationship to be an ongoing one. There are a risk-neutral principal (or a firm) and potentially very large number of risk-neutral agents. Both the principal and the agents are infinitely lived. They interact with one another at dates $t = 0, 1, 2, ...$ All the players share the common discount factor, δ, where $\delta \in (0, 1)$.

At each date, there are four stages. At the beginning of each date t, the principal decides

on how many agents to employ and offers a compensation package if she decides to employ any. The compensation package to agent i at date t consists of two parts: one is the base wage (w_{it}) and the other is the bonus payment (b_{it}) conditional on variables that are observable by both the principal and the agents.[11] Hence the total payment (W_{it}) from the principal to the agent i is the sum of these two: $W_{it} = w_{it} + b_{it}$. I assume that once the principal hires N agents at date 0, she continues to contract with the same agents at later dates unless any of them decides to leave the firm.

After being offered a compensation package, each agent decides whether or not to accept the offer. Let $d_{it} \in \{0, 1\}$ denote agent i's decision on whether to accept ($d_{it} = 1$) or not ($d_{it} = 0$). If $d_t = \prod_i d_{it} = 0$, i.e., any of the N agents rejects the offer, then each agent gets his reservation utility, $\underline{u}_{it}$, and the principal also gets her reservation payoff, π_{0t}. I assume that reservation payoff for the principal is time-invariant, $\pi_{0t} = \underline{\pi}$, which is normalized to zero. Reservation payoffs to the agents are also assumed to be time-invariant and identical for all agents, $\underline{u}_{it} = \underline{u}$ such that $\underline{u} \geq 0$.

If $d_t = \prod_i d_{it} = 1$, then each agents makes his production effort, e_{it} at the third stage. The production effort of an agent i at date t, e_{it}, is neither verifiable in the courts nor observable to the principal. Hence a contract cannot be contingent on agents' performance levels.

I assume that production occurs if and only if all the agents who are offered contracts from the principal accept the offers. This assumption is without loss of both generality and optimality. It may be the case that only a proper subset of agents accept the contract offered by the principal and the principal is better off producing than getting reservation payoff, $\underline{\pi}$. This contract, however, can be replaced by a contract that is offered and accepted by all the agents.[12]

At the last stage, conditional on $e_t = (e_{1t}, e_{2t}, ..., e_{Nt})$ exerted by the agents, final output Y_{Nt} is realized and commonly observed by the principal and the agents, where Y_{Nt} is a noisy signal on $e_t = (e_{1t}, e_{2t}, ..., e_{Nt})$. Though Y_{Nt} is observable to all parties, this is assumed to be not verifiable in the courts.[13] Hence, a contract cannot be contingent on Y_{Nt}. To provide incentives, it is necessary to rely on bonus payments that are self-enforcing, not

[11] i in subscript denotes the agent and t the date.

[12] Moreover, this is renegotiation-proof since, as will be seen in the following, the optimal stationary contract requires that every agent accepts the offer.

[13] If all public information is contractible, then a sequence of short-term contracts may replace a long-term contract. See Fudenberg et al. (1990).

court-enforced.

In addition to the production effort (e_{it}), each agent makes a monitoring effort (m_{it}). I assume that each agent has a unit time endowment so that he chooses e_{it} and m_{it} such that $e_{it} + m_{it} \in [0, 1]$. Exerting effort ($e_{it}$, m_{it}) at time t costs the agent i $c_i = c_i(e_{it}, m_{it})$. As in the standard models, the cost of effort is assumed to be strictly increasing and convex in both e_{it} and m_{it} with $c_i(0, 0) = 0$.

Each agent may or may not observe the others' effort choice perfectly, after making monitoring effort. Monitoring effort also may or may not be observable to the principal. In the case that each agent can observe the others' effort choice perfectly and monitoring effort is observable to the principal, the agents report their observations to the principal. When each agent makes monitoring and production efforts induced by a contract and every report is consistent with each other, the principal pays out the bonus b_{it} to each agent i.[14] In the case that each agent imperfectly observes the others' choice of effort but monitoring effort is observable to the principal, the bonus payment is contingent on the reports of the signal of each agent's production effort.

I assume that payoffs to the agents are quasi-linear. Given quasilinearity combined with the risk-neutrality of the players, the normalized discounted continuation payoffs to the principal and the agents at time t are:[15]

$$\Pi_t = (1-\delta)\mathbf{E}_t \sum_{\tau=t}^{\infty} \delta^{\tau-t}\pi_\tau \text{ where } \pi_\tau = \mathbf{E}_\tau\{d_\tau(Y_\tau - \sum_{i=1}^{N} W_{i\tau}) + (1-d_\tau)\underline{\pi}\}$$

$$U_{it} = (1-\delta)\mathbf{E}_t \sum_{\tau=t}^{\infty} \delta^{\tau-t}u_{i\tau} \text{ where } u_{i\tau} = \mathbf{E}_\tau\{d_\tau(W_{i\tau} - C_{i\tau}) + (1-d_\tau)\underline{u}\}$$

where $C_{i\tau}$ is cost to the agent i at date τ.

3.2 Preliminary Analysis

Before I examine the detailed model, I present the preliminary results on the relational contracts. In a two-period model, Bull (1987) derives incentive constraints ensuring the firm pay promised bonuses that are not enforceable by law. MacLeod and Malcomson (1989, 1993) extend Bull's analysis to an infinite horizon game but consider only the symmetric informa-

[14] If there is only one or two agents, peer monitoring raises the issue of truth-telling. I assume that there are at least 3 agents.

[15] I use capital letters for the flow payoffs and small letters for the per-period payoffs.

tion case. My model is about the asymmetric information case with infinite interactions. Moreover, since I consider the repeated principal-multiagent model, the contract space is potentially very large and it is cumbersome to tackle the problem for all possible contracts. However, I can restrict my attention to a stationary contract due to lemmata presented in this section. The results on the relational contracts are straightforward extensions of the results on the one principal-one agent relational contract in Levin (2002b).

Optimality of Stationary Contract

The principal and the hired agents observe the entire history of realized total output, and other common observables. Depending on whether monitoring effort is observable to the principal and whether monitoring is perfect, common observables other than realized output and wage may consist of monitoring efforts of the agents, production efforts of the agents, or signals about production efforts. Among the observables, only compensation from the principal to the agents is verifiable in courts. As a result, any incentive provisions conditional on the commonly observable variables can be credibly enforced only through the outcomes of repeated interactions.

Over the course of repeated interaction, the players' strategy depends on the commonly observed history, h_t, available at the beginning of each period t, where h_t consists of common observables. Let Ψ_t be the set of common observables at period t other than the participation decisions, compensation package and realized output, i.e., $\Psi_t \subset \{e_{it}, m_{it}, s_{it}\}_i$. Then

$$h_t = (\{d_{i0}, W_{i0}, Y_0, \Psi_0\}_i, ..., \{d_{it-1}, W_{it-1}, Y_{t-1}, \Psi_{t-1}\}_i).$$

An action profile for the principal at t is $\{W_{it}\}_i$ and action profiles for the agents are

$$\{d_{it}, e_{it}, m_{it}\}_i$$

which are mappings from history to some action in the set of actions.[16]

A strategy for each player is a collection of action profiles for each t. A relational contract is the one describes (i) the compensation package, (ii) the agents' participation decisions, and (iii) each agent's action, given that contract is accepted. A relational contract is self-enforcing if the strategy profile is a Nash equilibrium following every history. In other words, a contract

[16]By requiring each player's actions at t to be contingent on common observables, I focus on perfect public equilibria (Fudenberg, Levine and Maskin (1994)).

is self-enforcing if agents' individual rationality and incentive compatibility constraints are satisfied and the bonus payments are credible.

The set of self-enforcing relational contracts is potentially very large. However, as Levin (2002b) shows, I can simplify the analysis in a way that the problems of efficient production and of the distribution of the surplus are separately examined. Let S denote the surplus generated within a relationship among the principal and the agents. Also let $\underline{S} = \underline{\pi} + \sum_{i=1}^{N} \underline{u_i} = \underline{\pi} + N\underline{u}$ denote the sum of the reservation payoffs. The first lemma states that the joint surplus created by any self-enforcing contract can be divided arbitrarily among the players, provided that their individual rationality constraints are satisfied.

Lemma 1 *Suppose some self-enforcing relational contract generates total surplus $S/(1-\delta)$ larger than the reservation surplus $\delta\underline{S}/(1-\delta)$ (from the view point of $t=0$). Then for any $\lambda \in [0,1]^{N+1}$ with $\sum_{i=0}^{i=N} \lambda_i = 1$, there exits a self-enforcing relational contract that generates surplus $S/(1-\delta)$ and payoffs $\tilde{u_i}^0 = \underline{u_i} + \lambda_i(S - \delta\underline{S})/(1-\delta)$.*

The lemma says that as long as enough surplus is generated within relationships, there exists a self-enforcing contract that achieves the given surplus. By lemma 1, I can divide the problem into two parts: joint surplus maximization and the distribution of the surplus. As a result, I set aside the issue of surplus distribution and focus on the maximization of joint surplus as the objective of contract design.

Among all possible contracts in the set of self-enforcing contracts, this paper restricts attention, without loss of generality and optimality, to 'stationary' self-enforcing contracts as in Levin (2002b). Before showing this, define the stationary relational contract:

Definition 1 *A relational contract is stationary if at every t on the equilibrium path, $W_{it} = w_i + b_i$, $e_{it} = e_i$ and $m_{it} = m_i$ for some $w_i \in \mathbb{R}$, $b_i : \Phi \to \mathbb{R}$ and $e_i + m_i \in [0,1]$, where Φ is the set of common observables.*

By definition, a stationary relational contract is time-invariant in that the initial wage scheme chosen (and offered) applies to all dates t. Restricting attention to the set of stationary contracts is plausible in a situation where the principal does not have the power to make long-term commitment. Moreover, given the stationary contract offered, if one of the players fails to follow the contract, then separation occurs and each player gets her/his reservation payoff. Since every agent accepts the contract offer from the principal in equilibrium, this does not entail any loss. Lemma 2 shows that confining attention to stationary relational

contracts simplifies the analysis a lot without loss of optimality.

Lemma 2 *If there is an optimal relational contract, then there exist an optimal stationary relational contract.*

As a result of lemma 1 and 2, I restrict my attention to the stationary contacts. In the following, I drop the subscript denoting time unless it causes confusion.

Among stationary contracts, I look at a multilateral stationary contract rather than a sequence of bilateral contracts. That is, as mentioned above, production occurs if and only if all the agents accept the principal's offer. If I consider bilateral contracts, then the analysis becomes complicated without gaining much since with a bilateral contract, I need to solve the optimal contracting problem recursively. Generally, the multilateral contract generates a larger surplus by equalizing the rates of marginal benefit and marginal cost of performance across relationships. So my attention to multilateral contract comes without loss of optimality.[17] [18]

4 Detailed Model

In this section, I present a detailed model incorporating gains from specialization and monitoring costs. Based on the results in the previous section, I restrict my attention to the multilateral stationary contract and formulate the problem that the principal solves. Then I present my results on the bounded firm size under perfect monitoring with observable monitoring effort.

4.1 Technology

To incorporate the gains from specialization, it is necessary to consider more than one task since there is no gain from the division of labor with only one task. I suppose a continuum of tasks with perfect divisibility. Let the space of tasks be $(T, \mathcal{B}, \mu)$, where $T = [0, 1]$ is the set of tasks, $\mathcal{B}$ is the Borel σ-algebra on T and μ is the Lebesque measure. The principal has an agent i perform a set of tasks T_i such that $T_i \subset T$. I assume that every task must

[17] For more detailed comparison between multilateral and bilateral contracts, see Levin (2002a).

[18] In addition, the agents are on the short side in my model. If the agents compete one another *à la* Bertrand, then every agent ends up with the same contract in equilibrium.

be performed so that $\bigcup_i T_i = T$.[19] Since every agent is identical and each task is equally difficult, I restrict my attention to the symmetric assignment of tasks. Then I can express the size of the tasks for each agent in a unique way. Let A be the number of repetitions of all tasks, T. The size of each agent's set of tasks is then given by $\mu(T_i) = \frac{A}{N}$ for all i where $A \in [1, N]$. Different task assignments reflect different modes of organization. For example, if $A = 1$, then $\mu(T_i) = \frac{1}{N}$ and each agent performs different set of tasks. If $A = N$, then $\mu(T_i) = \frac{N}{N} = 1$ and each agent performs the same set of tasks, $T_i = T = [0, 1]$.

Given the task assignment, T_i, and $\mu(T_i) = \frac{A}{N}$, when exerting effort level e_i, each agent i produces intermediate products according to the following technology:

$$y_i = \frac{1}{\mu(T_i)} e_i = \frac{N}{A} e_i$$

Note that gains from specialization are embedded in this specification. Given the number of agents hired (N) and the level of effort (e_i), individual production, y_i is decreasing in the repetitions of all tasks (A). That is, the smaller A is and the smaller the set of tasks given to an agent i, the higher individual productivity is. This specification reflects Adam Smith's idea that higher individual productivity is achieved, the more specialized the task assignment is. Since individual productivity depends on the task assignment, how to assign tasks to the agents is a part of organization design, as will be seen in a later section.

The final output to the principal is defined as a function of all inputs produced by the agents:

$$Y_N = f(y_1, y_2, ..., y_N) + \epsilon$$

where ϵ is a random shock. Assume $\mathbf{E}[\epsilon | e = \{e_1, e_2, ..., e_N\}] = 0$, and ϵ is i.i.d across time. It is also assumed that the random shock affects final output after intermediate products are aggregated but before the principal observes aggregate production.

I make the following assumptions about the final output:

Assumption T1: Given N, f is linearly homogeneous, increasing and concave with $f(0, ..., 0) = 0$.

Assumption T2: Given N, f is symmetric in its arguments and $\frac{\partial^2 f}{\partial y_i \partial y_j} \geq 0$.

[19] One may wonder whether the final output may be produced even if only some proper subset of tasks is performed. I can model this situation by considering quality of output in a way that quality is a function of the size of tasks performed. As long as the quality is proportional to the size of the tasks performed, this does not change the results.

Note that T1 and T2 are about the properties of the production function with a given number of agents. T1 is a standard assumption implying that the final output production function exhibits constant returns to scale. Concavity is imposed for the maximization problem to be well-behaved for given N. T2 is technological interdependence. That is, complementarity and task specialization are considered together.

In addition, I impose a condition on how maximum average productivity varies as the number of agents changes:

Assumption T3: $f(\mathcal{I}_N) \leq N$ where $\mathcal{I}_N = \underbrace{(1, ..., 1)}_{N \ 1s}$.

Assumption T3 implies that average productivity is bounded by the number of agents. To see this, by assumption T1,

$$f(y_1, y_2, ..., y_N) = f(\frac{N}{A}e_1, ..., \frac{N}{A}e_N) = \frac{N}{A}f(e_1, ..., e_N)$$

Hence the average productivity is

$$\frac{f}{N} = \frac{1}{A}f(e_1, ..., e_N)$$

By the monotonicity from T1, this has a maximum when $e_i = 1$ for all i and $A = 1$. What T3 means is that maximum average productivity is asymptotically bounded by N. Roughly speaking, T3 allows average productivity to exhibit local increasing returns to scale but not global increasing returns to scale with respect to N. In other words, T3 excludes only the case in which per-capita productivity globally increases at increasing rates. In my view, it is realistic to assume that as N gets large, non-increasing returns of average productivity prevail and, as a result, average productivity is asymptotically bounded. Note, however, that I do not assume away increasing returns to scale of total productivity. Even though T3 imposes a bound on average productivity, this does include the case where the total output function exhibits increasing returns to scale with respect to N. And this covers fairly general class of production functions. For example, if f is Cobb-Douglas or Leontief, then $f(\mathcal{I}_N) = 1$ for all N, while in case of the linear technology, $f(\mathcal{I}_N) = N$.

I make one more assumption that implies non-decreasing average productivity.

Assumption T4: $f(\mathcal{I}_N) \leq f(\mathcal{I}_{N'})$ for all $N' > N$.

4.2 Cost of Effort

The agents exert two types of effort. One is production effort (e_i) and the other is peer monitoring effort (m_i). When making effort (e_i, m_i), the agent i incurs a cost given by $c_i = c_i(e_i, m_i)$, which is convex and strictly increasing.

It is assumed that it is too costly for the principal to monitor the agents and that the principal cannot discern from her observation of final output the individual contributions because of informational and technological externalities. Instead, this paper supposes that the agents perform monitoring activities and report their observations of signals to the principal.

This section considers the case of binary choice of monitoring effort, which is observable to the principal. I assume that the agents can observe true effort levels of the others only when the required monitoring effort is exerted. I think of the (required) monitoring effort as a function not only of the number of agents but also of task assignments. For monitoring effort (m_i), I impose the following assumption:

Assumption M1 $m_i(N', A) > m_i(N, A)$ for all $N' > N$ and all $A \in [1, N]$.

Assumption M1 says that monitoring effort is increasing in the number of coworkers, given the task assignment. The more an agent monitors, the higher monitoring effort is required for observing others' effort.[20]

Assumption M2 $m_i(N, A') > m_i(N, A)$ for all $A > A'$ and all N.

M2 implies that given the number of agents, the more overlap in the task assignments there is, the lower the required monitoring effort. Presumably, it is much easier to monitor those who work on the same types of tasks than those who work on tasks that one is not familiar with.

In addition to the cost of effort, an agent incurs another kind of cost when performing tasks with size $\mu(T_i) = \frac{A}{N}$. I call this adaptation cost, which is defined as $\mu(T_i) \cdot K = \frac{A}{N} \cdot K$ where K is a positive constant. This adaptation cost reflects the idea that the division of labor makes it possible to save fixed cost, say time, which is lost in passing from one type of work to another.

[20] One might wonder whether an agent has to monitor all the other agents. This, however, needs not be the case. For example, suppose an agent monitors some group of agents. Then M1 implies that as N gets large, it becomes harder to disentangle individual contributions and to observe true effort levels.

In sum, total cost that an agent incurs is[21]

$$C_i = c_i(e_i, m_i) + \frac{A}{N}K$$

I assume that all the agents have the same cost function so that $c_i(\cdot,\cdot) = c(\cdot,\cdot)$ is strictly increasing and convex.

4.3 Benchmark Cases

One Period Problem

Suppose that the principal and the agents interact in only one period. Since there is no future in a single period interaction, the principal has no incentive to pay bonuses after the benefit to the principal is realized. The agents, in turn, have no incentive to exert any production or monitoring effort. Then again, the principal has no incentive to set a positive court-enforced wage. So the equilibrium in one-period problem is $e_i = 0$, $m_i = 0$, $w_i = 0$ and $b_i = 0$ for all i.

First-Best

In the first-best, there is no need to monitor because there is no problem of asymmetric information. Clearly the optimal assignment of tasks is $A = 1$ and $\mu(T_i) = \frac{1}{N}$. That is, complete specialization is optimal in the first-best.

Considering these facts, define the first-best profit, $V^{FB}(N)$ as follows:[22]

$$\begin{aligned} V^{FB}(N) \quad &= \max_{\{e,w\}} E(Y_N \mid e) - \sum_{i=1}^{N} c(e_i) - AK - N\underline{u} \qquad (1) \\ s.t. \quad &(IR_A)\ \ w_i - c(e_i) - \frac{A}{N}K \geq \underline{u} \quad \forall i \in \{1, 2, ..., N\} \\ &(FE)\ \ 0 \leq e_i \leq 1 \quad \forall i \in \{1, 2, ..., N\} \end{aligned}$$

where (FE) denotes the feasibility of effort choices. Assume $V^{FB}(1) > 0$. This implies $V^{FB}(N+1) > V^{FB}(N)$ for all N and hence the size of the firm is unbounded in the first-best. Also assume that $\frac{\partial E(Y_N|e)}{\partial e_i}$ is non-decreasing in N so that the first-best effort level, e_i^{FB},

[21] One might think of fixed cost, $\frac{A}{N}K$, as learning cost that is paid only at date 0. The results do not change with this interpretation.

[22] I abuse notation by denoting cost of effort, $c(e)$, which is assumed to be strictly convex and increasing.

is non-decreasing in N for all i.

4.4 Multilateral Relational Contracts

In this section, I consider the case of perfect monitoring with observable monitoring effort. I will discuss how the results under perfect monitoring can be extended to the case of imperfect monitoring in a later section.

For the contract to be implementable, the following constraints should be satisfied. The first constraint is individual rationality constraint for each agent:

$$(IR_A) \qquad w_i + b_i - c(e_i, m_i) - \frac{A}{N}K \geq \underline{u} \qquad \forall i \in \{1, 2, ..., N\}$$

Unless the agent gets at least the reservation utility, he would not accept the contract offer.

The contract should also stipulate effort levels in an incentive compatible way. The incentive compatibility conditions for the agents are:

$$(IC_A) \qquad w_i + b_i - c(e_i, m_i) - \frac{A}{N}K \geq w_i - \frac{A}{N}K \qquad \forall i \in \{1, 2, ..., N\}$$

Note that with perfect monitoring, the incentive constraint collapses into a simple form. The LHS is the level of utility when the agent i makes both production and monitoring efforts. The RHS is the utility when he does not make any effort. If agent i does not exert any production effort, he knows that the other agents will monitor him and he will not be paid a bonus payment. Therefore, he has no incentive to perform any monitoring activity if he does not exert any production effort. When the agent i does not exert any production effort, his compensation is only the base wage and cost of effort consists of the adaptation cost only.

In addition to incentive compatibility for the agents, the principal should have incentives to pay out the bonus. Otherwise, the bonus payments are not credible and no incentives would be provided to the agents. The condition for a credible bonus payment, or the principal's incentive compatibility constraint is:

$$(IC_F) \qquad \frac{\delta}{1-\delta}(\pi - \underline{\pi}) \geq \sum_{i=1}^{N} b_i$$

The LHS is the discounted expected value of profits that the principal can get in an on-going relationship while the RHS is the sum of the bonus payments. If this constraint is violated,

the principal is better off not paying out the bonus payments. That is, the bonus payments are not credible and the effort levels dictated by the contract cannot be implemented. In my model, the principal gets all the surplus. With a normalization of $\underline{\pi} = 0$, (IC_F) becomes

$$\frac{\delta}{1-\delta} S \geq \sum_{i=1}^{N} b_i$$

where S is the total surplus generated.

Given the above set of constraints, the principal chooses the number of agents (N), task assignment (A) (hence $\frac{A}{N}$), base wage (w_i), and bonus payment (b_i) so as to make the agents choose effort levels to maximize profit.

I solve the principal's problem in two steps. In the first step, given the number of the agents (N), I solve out for task assignment (A), effort level (e), base wage (w) and bonus (b). In principle, I have maximum profit as a function of the number of agents, N. Then I choose the optimal number of agents, N, by examining this value function.

First, consider the problem with the given number of agents hired, N. Given the number of agents hired, the principal's problem is similar to the standard one except the fact that I consider the task assignment (A) and costly monitoring. An optimal stationary relational

contract solves the following:[23]

$$
\begin{aligned}
V^{SB}(N) \quad &= \max_{\{A,e,w,b\}} E(Y_N|e) - \sum_{i=1}^{N} c(e_i, m_i) - AK - N\underline{u} && (2)\\
s.t. \quad & (IR_A)\ \ w_i + b_i - c(e_i, m_i) - \frac{A}{N}K \geq \underline{u} && \forall i \in \{1, 2, ..., N\}\\
& (IC_A)\ \ w_i + b_i - c(e_i, m_i) - \frac{A}{N}K \geq w_i - \frac{A}{N}K && \forall i \in \{1, 2, ..., N\}\\
& (IC_F)\ \ \sum_{i=1}^{N} b_i \leq \frac{\delta}{1-\delta}(E(Y_N|e) - \sum_{i=1}^{N} c(e_i, m_i) - AK - N\underline{u})\\
& (FE)\ \ 0 \leq e_i + m_i \leq 1\ \ \forall i \in \{1, 2, ..., N\}\\
& (TA)\ \ A \in [1, N]
\end{aligned}
$$

The last two constraints are the feasibility of effort and the task assignment. Given the number of agents, N, and the task assignment, A, this program is well-defined. That is, by assumption T1, the problem is concave in $e = (e_1, ..., e_N)$ and the set of constraints are compact. Therefore, there exists a maximizer e^*. The problem is continuous in A that is chosen from a compact set. So there exists an optimal choice of A. In the following, I look at a symmetric equilibrium in which every agent exerts the same amount of production and monitoring efforts and the principal offers an identical contract to all the agents hired.

4.5 Bounded firm size

In the first-best, the size of the firm is unbounded. Given assumption T1-T4, the larger the size of the firm, the higher the average productivity. Moreover, the profit to the principal is increasing in the number of the agents. In the following, I show that firm size is bounded in the second best.

I make one additional assumption on monitoring effort:

[23]With this formulation, it is implicitly assumed that organizing production without monitoring is not profitable. This can be justified as follows. Let $G(Y_N \mid e = \{e_1, e_2, ..., e_N\})$ be the distribution function of final output conditional on the production effort choices of the agents. If $\frac{\partial G}{\partial e_i}$ goes to zero sufficiently fast as N increases, then it is never profitable to hire the large number of agents. Or alternatively, the variance of a signal on production effort is infinite unless required monitoring effort is made. As a result, only fixed wage contract is feasible in the absence of any monitoring. Though I do not explicitly model the informativeness of monitoring, I assume that monitoring generate informative enough signals so that it is always profitable to induce monitoring effort.

Assumption M3 $\frac{dm_i(N,A(N))}{dN} > \frac{\theta-1}{N^\theta}$ for all $N \geq 1$ with $m_i(1,1) = 0$ and $2 < \theta < \infty$.[24]

M3 concerns the rate of change of the required monitoring effort. Though it requires monitoring effort to be increasing in the number of agents, it does not impose the condition that monitoring technology exhibits decreasing returns to scale. If the monitoring technology exhibits decreasing returns to scale, then monitoring effort is a convex function of the number of agents, which is a less interesting case. What M1 and M3 together impose is that monitoring effort is an increasing function of the number of agents and the rate of increase is bounded below by some number that goes to zero as N gets large.

When the first-best is not implementable, the first two sets of constraints, $((IR_A), (IC_A))$ are binding in equilibrium. Then the set of constraints on individual rationality and incentive compatibility in the above problem collapses into the following simple expression:

$$w_i = \frac{A}{N}K + \underline{u}, \quad b_i = c(e_i, m_i)$$

Combining these with (IC_F) becomes

$$(RC) \qquad \sum_{i=1}^{N} c(e_i, m_i) \leq \frac{\delta}{1-\delta}(E(Y_N \mid e) - \sum_{i=1}^{N} c(e_i, m_i) - AK - N\underline{u})$$

(RC) is a constraint that should be satisfied for any contract to be implementable.

Given the above constraint, I begin the discussion with a lemma on the choices of production effort in the second-best.

Lemma 3 *$e^{SB} \leq e^{FB}$ for all N.*[25]

Lemma 3 says that the production effort levels in the second-best are never greater than those in the first-best. In my model, there are two distortions. One is from the moral hazard problem, which is a standard distortion. The other is the requirement for monitoring. Moral hazard can be mitigated through monitoring. But the requirement for monitoring is increasing and may crowd out the production effort. As a result, the second-best effort levels are less than the first-best ones.

Armed with lemma 3, I use (RC) to show the limit on firm size. If (RC) is violated,

[24] I treat the number of agents, N, as a continuous variable to avoid the technical difficulty of solving for an optimal integer. One justification could be that the principal may hire one part-time agent.

[25] Since I look at the symmetric equilibrium, I drop subscripts denoting the agents.

then there does not exist a relational contract that supports a surplus, regardless of the size of the surplus. For δ sufficiently close to zero, (RC) cannot be satisfied, but this is not an interesting case. I want to show that regardless of δ, (RC) is violated for large N. Indeed, under the assumptions I have made, this is the case.[26]

Proposition 1 *Under assumptions T1-T4, and M1-M3, the size of the firm is bounded in the second-best.*

The driving force of the result is the difficulty of monitoring and associated increases in monitoring cost. As the number of agents increases, the required level of monitoring activity increases and crowds out the production effort. As a result, the large potential gains from specialization cannot be achieved with the large number of agents. Two corollaries follow immediately.

Corollary 1 *The size of the firm is inversely related to monitoring technology, θ.*

Corollary 2 *Consider two different production functions, f and g such that $f(\mathcal{I}_N) < g(\mathcal{I}_N)$ for all N. Then the size of the firm under g is no less than that under f.*

The first corollary states that given production technology, the size of the firm is inversely related to the efficiency of monitoring technology. The second corollary says that the higher the gains from specialization, the larger the firm size. Though it is quite straightforward, this does have an important implication on firm size. That is, differences in monitoring difficulty and the magnitude of gains from specialization result in different firm sizes, otherwise identical. These are a simple verification of Coase's (1937) argument on the relationship between cost of organization and firm size.[27] The results of corollaries are also in accordance with casual observations that products subject to great variability are produced by small firms while standardized products are made by large firms.

[26] One might wonder, if a small size firm can make profit, can a large firm replicate the small firm and make infinite profit? This replication may be possible if a large firm consisting of divisions can condition compensation *explicitly and exclusively* on each division's contribution, which, I think, is rare. Moreover, in my model, the final output depends on the inputs of all agents. Consequently, breaking a large firm into small firms does not enable the large firm to replicate and make infinite amounts of profit.

[27] Coase (1937) says, "other things being equal, a firm will tend to be larger, the less the costs for organizing and the slower these costs rise with an increase in the transactions organized."

4.6 An Example with CES Production Function

As in the standard hidden action model, the predictions are not sharp unless I consider specific functional forms. Other than the bounded firm size, it is hard to explore the properties of other variables. In this section, I consider a specific functional form, the CES production function. The task space is the same as in the previous section. Let $T = [0, 1]$ be the set of tasks. An agent i performs a set of tasks $T_i \subset T$ with size $\mu(T_i) = \frac{A}{N}$ for all i where $A \in [1, N]$.

The final output is given by the CES function

$$Y_N = (\sum_{i=1}^{N} y_i^{\rho})^{1/\rho} + \epsilon, \quad \rho \leq 1$$

where y_i is the production of intermediate output by agent i and $E[\epsilon \mid e = \{e_1, e_2, ..., e_N\}] = 0$.

Given the task assignment, T_i with $\mu(T_i) = \frac{A}{N}$, when exerting effort level e_i, each individual i produces

$$y_i = \frac{1}{\mu(T_i)} e_i = \frac{N}{A} e_i$$

The aggregate function, then, becomes

$$Y_N = \frac{N}{A} (\sum_{i=1}^{N} e_i^{\rho})^{1/\rho} + \epsilon$$

and

$$\mathbf{E}[Y_N \mid e = \{e_1, e_2, ..., e_N\}] = \frac{N}{A} (\sum_{i=1}^{N} e_i^{\rho})^{1/\rho}$$

The assumptions T1-T4 are satisfied for $\rho = 1, 0$ and $-\infty$ that correspond to linear, Cobb-Douglas and Leontief technology respectively.[28]

Costs incurred by agent i when performing a set of tasks, T_i, exerting effort, e_i, and performing monitoring activities, m_i, are assumed to take the following form:

$$C_i = \frac{e_i^2}{2} + \frac{m_i^2}{2} + \frac{A}{N} K$$

[28] Even though only three values of ρ satisfy assumption T4, the assumption is not so restrictive as one might think. Even if I drop assumption T, firm size is bounded with appropriately chosen θ.

where K is a positive constant. Furthermore, I consider a specific monitoring technology:

$$m_i = 1 - \frac{1}{(2N - A)^\theta}$$

where θ measures the efficiency of the monitoring.

Given restrictions on the production function and cost function, the first-best production effort is $e^{FB} = 1$ for all three cases and $V^{FB}(N) = \frac{N}{2}$ for Cobb-Douglas or Leontief technology and $V^{FB}(N) = \frac{N^2}{2}$ for linear technology. Therefore, the larger the firm is, the higher the profit is in the first best. However, there exists a bound on firm size in the second best.

Corollary 3 *If $\theta > 0$, then the size of the firm is bounded for Cobb-Douglas or Leontief technology. If $\theta > 1$, then the size of the firm is bounded for linear technology.*

Corollary 4 *e is decreasing in N.*

Corollary 5 *A could be greater than 1.*

Corollary 3 is a result of straightforward calculation. Corollary 4 says that production effort is monotonically decreasing in the number of agents. Corollary 5 has an interesting implication. Depending on the relative magnitude of monitoring costs and gains from specialization, it may be the case that the task assignment may overlap among the agents. When monitoring costs increases fast, complete specialization may not be achieved within a firm. On the other hand, when the gains from specialization is more important than increases in monitoring costs, then complete specialization is a better choice for the principal and intrafirm division of labor may not be limited by agency cost.

4.7 Discussion of Imperfect Monitoring

When agents cannot observe the true effort levels of the others, a further complication arises. In this case, the principal receives only noisy signals on the production effort choices of the agents. Tailoring incentives is difficult and inducing the desired levels of effort is more costly. Therefore, when the agents observe some noisy signals about others' production efforts but monitoring efforts are observable, one can easily extend the previous result.

Given a binary choice of monitoring, the principal cannot do better under imperfect monitoring than under perfect monitoring. Hence I have

Proposition 2 *Suppose monitoring effort is observable to the principal but the agents cannot observe the true effort levels. Under assumptions T1-T4 and M1-M3, the size of the firm is bounded in the second-best.*

I relegate a detailed discussion of imperfect monitoring to the next section.

5 Extensions

In the previous section, I considered the binary choice of monitoring effort. That is, only when the required monitoring effort is made, signals on production effort are available. This section relaxes this assumption and extends analysis to the case of continuous monitoring choice. I consider two cases. One is when monitoring effort is observable to the principal and the other is when monitoring effort as well as production effort is private information to the agents. In both cases, I assume that monitoring is imperfect, i.e., monitoring activities generate noisy signals on the production effort. In the following, I drop assumptions M1-M3 while maintaining assumptions T1-T4.

5.1 Imperfect Monitoring with Observable Monitoring Effort

Consider the case where agent i's production effort, e_i, is private information but it stochastically affects the publicly observed signal s_i among the agents. Suppose that s_i is distributed according to the twice continuously differentiable conditional cumulative function $H(s_i|e_i, m_{-i})$, which does not depend on e_{-i}. That is, conditional on e_i, each s_i is i.i.d. across players. s_i for all i is also assumes to be i.i.d. across time. Suppose further that aggregate monitoring effort affects the distribution, i.e., $m_{-i} = \sum_{j \neq i} m_j$. I consider the separable cost function in e_i and m_i, $c_i(e_i, m_i) = c(e_i) + \gamma(m_i)$, where $c(\cdot)$ and $\gamma(\cdot)$ are strictly increasing and convex with $c(0) = \gamma(0) = 0$, $c'(0) > 1$, $c(1) < \infty$, $\gamma'(0) > 1$, and $\gamma(1) < \infty$.

Let b_i denote the total bonus payment to agent i. b_i consists of two parts: the bonus payment conditional on the signal, $\beta_i(s_i)$, and the bonus payment conditional on monitoring effort, $\chi_i(m_i)$. Since monitoring effort is observable, $\chi_i(m_i) = \gamma_i(m_i)$.[29]

[29]If the bonus payment depends only on the signal realization, the agents has no incentive to monitor the others. Then this, in turn, goes back to the situation where the principal observes only final output without having no additional information.

Under imperfect monitoring, the incentive constraints for the agents become

$$(IC_A) \qquad e_i = \arg\max \; w_i + \mathbf{E}[\beta_i(s_i)|e_i, m_{-i}] + \chi_i(m_i) - c(e_i) - \gamma(m_i) - \frac{A}{N}K \qquad \forall i$$

In order to apply the first order approach, I make the following assumptions on the distribution of the signals in the spirit of Jewitt (1988):

Assumption H1 The likelihood ratio $\frac{h_i(s_i|e_i,m_{-i})}{h(s_i|e_i,m_{-i})}$ is increasing and concave in s_i for all e_i, where $h(s_i|e_i, m_{-i})$ is density of H and $h_i(s_i|e_i, m_{-i}) = \frac{\partial h(s_i|e_i,m_{-i})}{\partial e_i}$.

Assumption H1 states that an increase in e_i stochastically increases the value of s_i in the sense of first-order stochastic dominance. This is also known as the monotone likelihood ratio property (MLRP). I make two more assumptions so that the first-order approach is valid.

Assumption H2 $\int_{-\infty}^{z} H(s_i|e_i, m_{-i})ds_i$ is nonincreasing and convex in e_i for each z.

Assumption H3 $\int s_i dH(s_i|e_i, m_{-i})$ is nondecreasing and concave in e_i.

H2 and H3 together imply that the agent cannot improve the distribution of output by randomizing the amount of effort in the sense of second-order stochastic dominance. Assumptions H1-H3 validate the use of the first-order approach (Jewitt (1988)).[30]

In addition, I assume H4 for the effect of monitoring effort on the distribution of the signal.

Assumption H4 $H(s_i|e_i, m_{-i})$ is a mean-preserving spread of $H(s_i|e_i, m'_{-i})$ if $m'_{-i} > m_{-i}$.

H4 implies that m_{-i} does not affect the mean but rather the variance of the signal s_i. In other words, only the precision of the signal is increasing in monitoring effort from others.[31]

By applying the first-order approach, (IC_A) can be replaced by the first-order condition:

$$\frac{\partial \mathbf{E}(\beta_i(s_i)|e_i, m_{-i})}{\partial e_i} - c'(e_i) = 0 \qquad \forall i$$

As Levin (2002a, Theorem 3) shows, the optimal bonus scheme takes an one-step form under assumptions H1-H3. That is, the bonus payment conditional on the signal takes one of two

[30]The assumptions H1-H3 are weaker than widely used Mirrlees-Rogerson condition for the validity of the first-order approach.

[31]Note that assumptions H1-H4 are about the signal distribution with a given number of agents.

values, depending on the cutoff signal $\hat{s}_i$:

$$\beta_i(s_i) = \begin{cases} \overline{\beta}_i = \sup_{s_i} \ \beta_i(s_i) & \text{if } s_i > \hat{s}_i \\ \underline{\beta}_i = \inf_{s_i} \ \beta_i(s_i) & \text{otherwise} \end{cases}$$

where $\hat{s}_i$ is defined by $h_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i}) = 0$ and $\{\hat{e}_i, \hat{m}_{-i}\}_i$ are effort levels stipulated by the contract. By the monotone likelihood ratio property (H1), $h_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i})$ is negative for $s_i < \hat{s}_i$ and positive for $s_i > \hat{s}_i$. Hence the bonus is monotone with respect to the signal's realization.[32]

With this one-step bonus scheme, the first-order condition becomes

$$\overline{\beta}_i - \underline{\beta}_i = -\frac{c'(\hat{e}_i)}{H_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i})}$$

where $H_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i}) = \int_{-\infty}^{\hat{s}_i} h_i(s_i|\hat{e}_i, \hat{m}_{-i})ds_i$. What matters in this expression is the difference between $\overline{\beta}_i$ and $\underline{\beta}_i$. Given the difference, the individual rationality constraints for the agents can be satisfied with an appropriately chosen base wage, w_i. Therefore, without loss of generality, I may assume the non-negativity of the bonus payment so that I can express the optimal amount of bonus as

$$\overline{\beta}_i = -\frac{c'(\hat{e}_i)}{H_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i})} \ , \ \ \underline{\beta}_i = 0$$

The maximum total bonus payment to agent i then becomes

$$\overline{b_i} = \overline{\beta}_i + \chi(\hat{m}_i) = -\frac{c'(\hat{e}_i)}{H_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i})} + \gamma(\hat{m}_i)$$

Under imperfect monitoring, the principal's ability to induce production effort by the agents is more limited than under perfect monitoring. This can be seen in the expression of the bonus payment. Under perfect monitoring, the principal only needs to pay out bonuses for production effort, which is equal to the cost of true production effort. With imperfect monitoring, the relevant constraints are the difference between maximum and minimum amount of bonus. As a result, the bonus payment for inducing a given level of effort is higher under imperfect monitoring.

[32]Rayo (2002) shows that statistically this is the uniformly most powerful test of Neyman-Pearson test of the null hypothesis $H_0 : e_i = \hat{e}_i$

For this bonus payment to be credible, the following condition should hold:

$$(RC) \qquad \sum_{i=1}^{N}\{\overline{\beta}_i+\gamma(\hat{m}_i)\} \leq \frac{\delta}{1-\delta}(E(Y_N|\hat{e})-\sum_{i=1}^{N}c(\hat{e}_i)-\sum_{i=1}^{N}\gamma(\hat{m}_i)-AK-N\underline{u})$$

As one can see from (RC), even in the case that the 'expected' amount of the total bonus payment is small relative to total surplus generated, (RC) constraint may not be satisfied. That is, the credibility of bonus takes care of the unlikely event in which every agent would receive the maximum amount of bonus, $\overline{\beta}_i$.

Unfortunately, not much is known about costly imperfect monitoring in a repeated game setting. So, to tie my hands, I consider a specific distribution function, the normal distribution.[33] Given monitoring effort, $m_{-i} = \sum_{j\neq i} m_j$, the signal on e_i, is distributed:

$$s_i \sim \mathrm{N}(e_i, \frac{(2N-A)^\theta \sigma^2}{\sum_{j\neq i} m_j}), \quad \theta > 2 \tag{3}$$

where σ is a positive constant. s_i is assumed to be i.i.d across the agents and across time. This specification imposes the conditions on the distribution of signals similar to assumptions M1-M3. First, the mean of the signal depends only on the true effort. Second, the variance is increasing in the number of agents (N), and decreasing in the task-overlap (A) and aggregate monitoring effort of agents other than agent i, (m_{-i}). That is, other things being equal, an increase in N makes the signal noisier. Increases in the overlap of the tasks among the agents and in the monitoring effort improve the precision of the signal.

Given the normal distribution, the cutoff level of signal is determined by

$$h_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i}) = 0 \;\; \Leftrightarrow \;\; \hat{s}_i = \hat{e}_i.$$

Therefore,

$$-H_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i}) = \sqrt{\frac{2}{\pi}}\frac{\sum_{j\neq i} m_j}{(2N-A)^\theta \sigma^2}$$

[33] To my knowledge, the normal distribution is the only one that satisfies assumptions H1-H4. Even though any distribution belonging to the exponential family satisfies assumptions H1-H3, only the normal distribution meets H4.

The maximum total bonus payment to agent i then becomes

$$b_i = \overline{\beta}_i + \chi(\hat{m}_i) = -\frac{c'(\hat{e}_i)}{H_i(\hat{s}_i|\hat{e}_i, \hat{m}_{-i})} + \gamma(\hat{m}_i) = \sqrt{\frac{\pi}{2}}\frac{(2N-A)^\theta \sigma^2}{\sum_{j\neq i}\hat{m}_j}c'(\hat{e}_i) + \gamma(\hat{m}_i)$$

Note that the bonus payment is decreasing in monitoring effort $\hat{m}_{-i}$.

Now, combining the above with (RC) gives

$$(RC) \quad \sum_{i=1}^{N}\left\{\sqrt{\frac{\pi}{2}}\frac{(2N-A)^\theta \sigma^2}{\sum_{j\neq i}\hat{m}_j}c'(\hat{e}_i) + \gamma(\hat{m}_i)\right\} \leq \frac{\delta}{1-\delta}(E(Y_N|\hat{e}) - \sum_{i=1}^{N}c(\hat{e}_i) - \sum_{i=1}^{N}\gamma(\hat{m}_i) - AK - N\underline{u})$$

As is mentioned above, the stringent condition that the maximum bonus payment should be taken into account for the credibility of the bonus is a binding constraint. As N increases, the level of bonus payment necessary to induce a given level of effort increases. As a result, it becomes more costly to induce production effort as firm size grows. Hence, the principal is rather better off inducing more monitoring effort and less production effort to control the size of the bonus payment. In the limit, the principal induces only an infinitesimal level of production effort, which is the following lemma.

Lemma 4 $\lim_{N\to\infty} e(N) = 0$.

In the absence of moral hazard, the production efforts are non-decreasing in the number of agents hired. However, due to the moral hazard problem, the principal should balance production effort against monitoring effort. If less monitoring effort is induced, then the principal needs to pay higher bonus payments to induce higher production effort. The bonus payments are increasing in the number of agents hired because the precision of the signal deteriorates. Hence, the principal is better off inducing higher monitoring effort and, in the limit, the production effort goes to zero.

Armed with the above lemma, the result on firm size is presented in case of imperfect monitoring but observable monitoring effort by the principal.

Proposition 3 *Suppose monitoring effort is observable but agents cannot observe the true effort level. Under assumptions T1-T4, and the given normal distribution of the signals, the size of the firm is bounded.*

Proposition 3 is in contrast with Holmström (1982, Theorem 3) stating that with the appropriately chosen group penalty scheme, the first-best can be approximated arbitrarily

closely. If Holmström's result applied to my model, firm size would be unbounded in the second-best since firm size is unbounded in the first-best due to increasing returns from specialization. In contrast to his result, firm size is limited in the second best. Even though the same distributional assumptions are made as Holmström, proposition 3 applies. The key differences are costly monitoring and the credibility of the bonus. If monitoring is not costly, then the second-best can be approximated arbitrarily close to the first-best even in the single period model. However, costly monitoring and the crowding effect of large numbers take the second-best away from the first-best and imposes an upper bound on firm size. In addition, Holmström assumes that if the principal has unbounded wealth, then the group penalty scheme can be replaced with a bonus scheme in which credibility is assumed. But once I take credibility of the bonus into account, I have a different result.

5.2 Imperfect Monitoring with Unobservable Monitoring Effort

If monitoring is imperfect and monitoring effort is unobservable, then a further complication arises because it is impossible to condition the contract on the unobservable monitoring effort. Hence, there is a moral hazard problem not only in the choice of production effort but also in that of monitoring effort.

In my model, the agents receive a payoff that is equal to their reservation payoff and are indifferent between staying and leaving. The agents' incentive to monitor is diluted under the base wage-bonus payment scheme presented in the previous sections since monitoring effort of agent i is costly but does not increase his payoff. Hence, each agent has less incentive to monitor the others when monitoring effort is not observable to the principal.

To proceed with the analysis on firm size, it is necessary to specify a signal distribution on monitoring effort, in addition to the signal distribution on production effort. I make similar assumptions on the distribution of the signals on monitoring effort as in the previous subsection. That is, I suppose that agent i's monitoring effort, m_i, stochastically affects the publicly observed signal ν_i among the agents. Suppose that ν_i is distributed according to the twice continuously differentiable conditional cumulative function $G(\nu_i|m_i)$, which does not depend on m_{-i}. I also assume that the first-order approach is valid with the distribution function, $G(\nu_i|m_i)$. That is,

(i) The likelihood ratio $\frac{g_i(\nu_i|m_i)}{g(\nu_i|m_i)}$ is increasing in ν_i for all m_i, where $g(\nu_i|m_i)$ is density of G and $g_i(\nu_i|m_i) = \frac{\partial g(\nu_i|m_i)}{\partial m_i}$.

(ii) $\int_{-\infty}^{z} G(\nu_i|m_i)d\nu_i$ is nonincreasing and convex in m_i for each z.
(iii) $\int \nu_i dG(\nu_i|m_i)$ is nondecreasing and concave in m_i.

As before, the bonus payment to agent i consists of two parts: one payment for production effort ($\beta_i(s_i)$) and the other for monitoring effort ($\chi_i(\nu_i)$).

The incentive constraints for the agents become

$$(IC_A)\ \ (\hat{e}_i, \hat{m}_i) = \arg\max_{\{e_i, m_i\}}\ \ w_i + \mathbf{E_h}[\beta_i(s_i)|e_i, m_{-i}] + \mathbf{E_g}[\chi_i(\nu_i)|m_i] - c(e_i) - \gamma(m_i) - \frac{A}{N}K \quad \forall i$$

where $\mathbf{E_h}$ and $\mathbf{E_g}$ denote the expectations taken with respect to H and G. Since the first-order approach is assumed to be valid, the incentive constraints can be replaced by first-order conditions. The first-order conditions are

$$\begin{aligned} \frac{\partial \mathbf{E}(b_i(s_i)|e_i, m_{-i})}{\partial e_i} - c'(e_i) &= 0 \\ \frac{\partial \mathbf{E}(\chi_i(\nu_i)|m_i)}{\partial e_i} - \gamma'(m_i) &= 0 \end{aligned}$$

Both of the bonus schemes for production and monitoring effort take an one-step form. Assuming the non-negativity of the bonus and the normal distribution for the signal given in the previous section, I have the following when the contract dictates $\{\hat{e}_i, \hat{m}_i\}$:

$$\overline{\beta}_i = \sqrt{\frac{\pi}{2}} \frac{(2N-A)^\theta \sigma^2}{\sum_{j\neq i} \hat{m}_j} c'(\hat{e}_i), \quad \underline{\beta}_i = 0 \tag{4}$$

$$\overline{\chi}_i = \frac{\gamma'(\hat{m}_i)}{-G_i(\hat{\nu}_i|\hat{m}_i)}, \quad \underline{\chi}_i = 0 \tag{5}$$

where $G_i(\hat{\nu}_i|\hat{m}_i) = \int_{-\infty}^{\hat{\nu}_i} g_i(\nu_i|\hat{m}_i)d\nu_i$ and $\hat{\nu}_i$ is defined by $g_i(\hat{\nu}_i|\hat{m}_i) = 0$.

Note that both $\overline{\beta}_i$ and $\overline{\chi}_i$ should be non-negative. Moreover, if $\overline{\beta}_i = 0$, then $e_i = 0$, and if $\overline{\chi}_i = 0$, $m_i = 0$. Neither $e_i = 0$ or $m_i = 0$ can be sustained as an equilibrium. Hence, both $\overline{\beta}_i$ and $\overline{\chi}_i$ should be positive.

Combining (4) and (5) with the credibility of bonus gives

$$(RC) \qquad \sum_{i=1}^{N} \{\overline{\beta}_i + \overline{\chi}_i\} \leq \frac{\delta}{1-\delta}(E(Y_N|\hat{e}) - \sum_{i=1}^{N} c(\hat{e}_i) - \sum_{i=1}^{N} \gamma(\hat{m}_i) - AK - N\underline{u})$$

It turns out that when N is large enough, the required bonus increases fast because the

moral hazard problem worsens and, as a result, the above (RC) cannot be satisfied. This implies that there exists a finite firm size that maximizes the principal's profit.

Proposition 4 *Suppose both production effort and monitoring effort are private information to the agents. Then under assumptions T1-T4, and the given distribution of the signals, the size of the firm is bounded.*

It is of no surprise that there exists a bound on firm size under imperfect monitoring with unobservable monitoring effort. Indeed, proposition 4 is a straightforward extension of proposition 3.

Remark 1: To induce higher production and monitoring effort, the principal may choose to share profit with agents. Under profit sharing, an agent has incentive to exert monitoring effort because higher monitoring effort increases the precision of the signal which induces higher production effort. In fact, the cost of inducing given levels of effort is lower under profit sharing. Nevertheless, firm size is bounded. Since the analysis is very similar to the above discussion, I relegate the discussion of profit-sharing to appendix B.

Remark 2: Incorporating hierarchical structure into the model does not change the results on the bounded firm size. Considers the principal hires M benevolent supervisors who have unit time endowment in each period and monitor N agents. Ignoring the integer problem, suppose that each supervisor monitors $\frac{N}{M}$ agents and hence spends $\frac{M}{N}$ time for each worker to be monitored. Given the information structure in this section, it can easily be shown that hierarchical structure does not change the results on firm size. If there exists moral hazard problem with the supervisors, then it becomes more costly to provide incentives to the supervisors and hence firm size is bounded. Moreover, as Itoh (1992) shows, in the extreme case where the supervisors collude with agents, the hierarchical structure has no value. In sum, the results on firm size is robust against introduction of hierarchy into the model.

6 Concluding Remarks

This paper studies the size of the firm under moral hazard by extending the principal-agent model. This paper explicitly takes into account benefits and costs that the principal faces

when hiring more than one agent. The benefits are gains from specialization. Hiring more agents makes it possible to achieve higher productivity through specialized task assignments. However, an increase in task specialization makes peer monitoring more difficult and costly. This paper shows that agency costs due to moral hazard are one factor that sets limits on firm size in a model where it would otherwise be unbounded.

In this paper, only monitoring costs are considered as a factor that sets bounds on firm size. Admittedly, there are other factors affecting the size of the firm, such as coordination costs, communication costs, and market size, to name a few. It is also implicitly assumed that there is no fixed factor such as capital so that the marginal product of labor does not decrease with the number of employees. Incorporating these factors into the analysis would enhance our understanding of the determinants of firm size.

In addition, there are several ways to extend this paper. First of all, this paper does not consider hierarchical structures. The firm is assumed to have a flat hierarchy in the sense that there is only one principal who controls all the agents hired. A natural question that arises is whether or not to hire specialists (supervisors) who specialize in monitoring. This introduces the hierarchical structure into my model. Incorporating hierarchy into the model would not change the main result on firm size. It is interesting, however, to examine how the hierarchical structure endogenously emerges and varies as firm size changes. Considering the size of a firm with an optimal hierarchy design will give a better understanding of internal structure of the firm.

Secondly, it is implicitly assume that the price of good is fixed at 1 and the firm is a price-taker. The firm can sell as much as possible in my model. It is likely, however, that the ability to expand firm size is limited by market conditions. Adding market considerations will complement this paper. An interesting related question is what a firm would do in response to changes in the market conditions such as price or the degree of competition. Through this analysis, one can better understand how firm size varies in response to changes in market conditions.

Another interesting avenue to go is to look at the problem of firm scope. Rather than looking at the size of the firm, one can ask how much scope the firm would choose, given the number of agents. This would make it possible to examine the division of labor across firms. These are important topics for future research.

Appendix A: Proofs of Results

Proof of Lemma1. Let $(\tilde{e}_i, \tilde{m}_i)_i$ be implemented by some self-enforcing contract,$(\tilde{w}_i, \tilde{b}_i)_i$. The supposition that the contract is self-enforcing implies the following three conditions are satisfied:

(a) $\mathbf{E}[\tilde{w}_i + \tilde{b}_i - C + \frac{\delta}{1-\delta}u] \geq \frac{1}{1-\delta}\underline{u}$

(b) $\{\tilde{e}_i, \tilde{m}_i\} = \arg\max_{\{e_i, m_i\}} \ \mathbf{E}[\tilde{w}_i + \tilde{b}_i - C + \frac{\delta}{1-\delta}u]$

(c) $\frac{\delta}{1-\delta}[\pi - \underline{\pi}] \geq \sum_{i=1}^{N} b_i$

Now consider the changes in w_i such that $\tilde{w}_i \neq w_i$ for $i = 1, ...N$. This has no incentive effect but changes the division of surplus between the principal and the agents. Hence choosing $\tilde{w}_i$ so as to satisfy individual rationality and incentive compatibility for the principal only changes the distribution. Therefore I can pick any $\lambda = (\lambda_0, \lambda_1, ..., \lambda_N)$ such that for $\sum_{i=0}^{i=N} \lambda_i = 1$, $\tilde{u}_i = \underline{u} + \lambda_i \frac{S - \delta\underline{S}}{1-\delta}$ for $i = 1, ..., N$ and $\tilde{\pi} = \underline{\pi} + \lambda_0 \frac{S - \delta\underline{S}}{1-\delta}$. ■

Proof of Lemma 2. Consider any optimal contract, $(\tilde{w}_i, \tilde{b}_i)_i$ that is not stationary. Let $\tilde{S}$ be the surplus generated and $(\tilde{e}_i, \tilde{m}_i)_i$ be the effort choices induced under this contract.

Since the original contract is self-enforcing,

(a) $\mathbf{E}[\tilde{w}_i + \tilde{b}_i - C + \frac{\delta}{1-\delta}u] \geq \frac{1}{1-\delta}\underline{u}$

(b) $\{\tilde{e}_i, \tilde{m}_i\} = \arg\max_{\{e_i, m_i\}} \ \mathbf{E}[\tilde{w}_i + \tilde{b}_i - C + \frac{\delta}{1-\delta}u]$

(c) $\frac{\delta}{1-\delta}[\pi - \underline{\pi}] \geq \sum_{i=1}^{N} b_i$

Let $\tilde{u}_i$ be the continuation payoff under the original contract.

Consider the following stationary contract $(\hat{w}_i, \hat{b}_i, \hat{e}_i, \hat{m}_i)_i$:

$$\hat{b}_i + \frac{\delta}{1-\delta}\hat{u}_i = \tilde{b}_i + \frac{\delta}{1-\delta}\tilde{u}_i, \quad \hat{w}_i = \hat{u}_i - E[\tilde{b}_i - c_i | e_i] \tag{6}$$

This yields the payoffs to the agents, $\tilde{u}_i$ for $i = 1, ..., N$. And the stationary contract satisfies the incentive constraints for the agents since (6) provides the same incentive powers to the agents. We can pick $\{\tilde{u}_i\}_{i=1}^{N}$ so that the participation constraints for the agents as well as for the principal are satisfied due to lemma 1. One can then repeat this stationary contract in the following periods so that a stationary contract gives the same payoffs.■

Proof of Lemma 3: Suppose on the contrary $e^{SB} > e^{FB}$. Then consider a new effort choice such that $\tilde{e} = (1-\epsilon)e^{SB} + \epsilon e^{FB}$ and $\tilde{m} = m^{SB}$. Then it follows that this is feasible and by concavity of the problem given N and A, this increases total profit. This contradicts to the optimality of the second-best contract. ■

Proof of Proposition 1: To show the boundedness of the firm size in the second best, I want to show that for sufficiently large N, there is no relational contract generating a surplus. In other words, I want to have (RC) violated for sufficient large N and thereafter. From (RC),

$$\sum_{i=1}^{N} c(e_i, m_i) \leq \frac{\delta}{1-\delta}(E(Y|e) - \sum_{i=1}^{N} c(e_i, m_i) - AK - N\underline{u})$$

$$(\Leftrightarrow) \quad \sum_{i=1}^{N} c(e_i, m_i) \leq \delta(E(Y|e) - AK - N\underline{u})$$

By assumption T1, the above can be expressed as:

$$\sum_{i=1}^{N} c(e_i, m_i) \leq \delta(\frac{N}{A}e(N)f(\mathcal{I}_{\mathcal{N}}) - AK - N\underline{u})$$

In the symmetric equilibrium,

$$c(e, m) \leq \delta(\frac{1}{A}e(N)f(\mathcal{I}_{\mathcal{N}}) - \frac{A}{N}K - \underline{u})$$

By assumption M3, $f(\mathcal{I}_{\mathcal{N}}) = o(\frac{1}{e(N)})$ and $f(\mathcal{I}_{\mathcal{N}})e(N)$ goes to zero as N goes to ∞. Hence the RHS goes to some non-positive number. On the other hand, the LHS goes to some positive number by assumption M3. Therefore, there exists $\overline{N}$ such that for all $N \geq \overline{N}$, (RC) cannot be satisfied. ■

Proof of Lemma 4: Note first that $e(N)$ is a bounded sequence. Therefore there exist lim sup and lim inf. Suppose that $\limsup_{N\to\infty} e(N) = e^{\top}$ where $e^{\top} > 0$. This implies that, for each ϵ, there there are infinite number of $e(N)$ such that $e^{\top} - \epsilon < e(N)$. Consider (RC)

$$\sum_{i=1}^{N}\{\sqrt{\frac{\pi}{2}}\frac{(2N-A)^{\theta}\sigma^2}{\sum_{j\neq i} m_j} + \gamma(m_i)\} \leq \frac{\delta}{1-\delta}(E(Y_N|e) - \sum_{i=1}^{N} c(e_i) - \sum_{i=1}^{N}\gamma(m_i) - AK - N\underline{u})$$

In symmetric equilibrium,

$$(1-\delta)\sqrt{\frac{\pi}{2}}\frac{(2N-A)^{\theta}\sigma^2}{(N-1)m} + \gamma(m)\} \leq \delta(\frac{1}{A}e(N)f(\mathcal{I}_{\mathcal{N}}) - \frac{A}{N}K - \underline{u})$$

But (RC) cannot be satisfied for $e(N)$ such that $e^{\top} - \epsilon < e(N)$ when N is large enough. Hence, $\limsup_{N\to\infty} e(N)$ cannot be positive. And $\liminf_{N\to\infty} e(N) \geq 0$. Therefore, $\lim_{N\to\infty} e(N) = 0$.

Proof of Proposition 3: This follows directly from lemma 4. ■
Proof of Proposition 4: The same as proposition 3, so omitted. ■

Appendix B: Bounded Firm Size under Profit Sharing

In this appendix, I show that even though the profit sharing scheme may induce agents' effort in production and monitoring at lower cost, firm size is bounded.

The principal shares her profit with a group of agents in a way that α of the profit is the share of the agents. And the agents share equally so that each agent's share of profit is $\frac{\alpha}{N}$. So it is like a sub-partnership among agents. The share of profit to the agents, α, could be a function of the number of agents. For the moment, I consider the case where α is fixed.[34] The result holds for any $\alpha \in (0, 1]$. Since the interest is not solving for the optimal sharing scheme but showing the existence of limit on firm size, this does not entail any loss of generality. Indeed the result does not change with consideration of unequal sharing rules among the agents.[35] Given the task assignment, the sub-partnership among the agents maximizes $\alpha\cdot$ profit.

If monitoring effort is observable to the agents, but not to the principal, then the analysis is identical to the case in section 5. So it is assumed that not only production effort but also monitoring effort are private information. After monitoring effort is exerted, a publicly observable signal on the effort level is available and according to the signal realization, the agreed bonus payments for production effort and monitoring effort are exchanged. The time line of the game is identical to the previous sections except for the fact that the bonus payments are decided by the (unanimous) agreement among agents.

To tie my hands, I make one additional assumption on technology:

Assumption T5

$$\frac{\partial}{\partial e_i} f(e_1, ..., e_N)|_{(e_1,...,e_N)=(e,...,e)} \leq 1 \quad \text{for all } e \in (0, 1]$$

[34]This can be justified by the assumption that the principal's bargaining power is fixed at $1 - \alpha$.
[35]Rayo (2002) characterizes the optimal sharing rule as well as compensation scheme in the team.

T5 implies that the average marginal productivity of e_i is bounded. That is, the average of the marginal productivity cannot grow at increasing rates as N increases. Note that this does not impose any bound on the marginal productivity of e_i. Total output is $\frac{N}{A}f(e_1, ..., e_N)$ and the marginal productivity of e_i is $\frac{N}{A}\frac{\partial}{\partial e_i}f(e_1, ..., e_N)$. Hence, what T5 imposes is that marginal productivity is bounded by $\frac{N}{A}$ that can be unbounded as N increases. Moreover, T5 is compatible with T1-T4. For example, Cobb-Douglas, Leontief and linear technologies satisfy T5 as well as T1-T4.

With profit sharing, the incentive constraint for each agent becomes

$$(IC_A)\ (\hat{e}_i, \hat{m}_i) = \arg\max_{\{e_i, m_i\}} \frac{\alpha}{N}\pi(N) + \mathbf{E_h}[\beta_i(s_i)|e_i, m_{-i}] + \mathbf{E_g}[\chi_i(\nu_i)|m_i] - c(e_i) - \gamma(m_i) - \frac{A}{N}K \quad \forall i$$

where $\mathbf{E_h}$ and $\mathbf{E_g}$ denote the expectations taken with respect to H and G. Since I assume the first-order approach is assumed to be valid, the incentive constraints can be replaced by first-order conditions. The first-order conditions are

$$\begin{aligned} \frac{\alpha}{N}\frac{\partial \pi}{\partial e_i} + \frac{\partial \mathbf{E}(b_i(s_i)|e_i, m_{-i})}{\partial e_i} - c'(e_i) &= 0 \\ \sum_{j \neq i} \frac{\alpha}{N}\frac{\partial \pi}{\partial e_j}\frac{\partial e_j}{\partial m_i} + \frac{\partial \mathbf{E}(\chi_i(\nu_i)|m_i)}{\partial e_i} - \gamma'(m_i) &= 0 \end{aligned}$$

Assuming the non-negativity of the bonus and the normally distributed signal as in section 5, I have

$$\overline{\beta}_i = \sqrt{\frac{\pi}{2}}\frac{(2N - A)^\theta \sigma^2}{\sum_{j \neq i} \hat{m}_j}[c'(\hat{e}_i) - \frac{\alpha}{N}\frac{\partial \pi}{\partial e_i}], \quad \underline{\beta}_i = 0 \tag{7}$$

$$\overline{\chi}_i = \frac{\{\gamma'(\hat{m}_i) - \frac{\alpha}{N}\sum_{j \neq i}\frac{\partial \pi}{\partial e_j}\frac{\partial e_j}{\partial m_i}\}}{-G_i(\hat{\nu}_i|\hat{m}_i)}, \quad \underline{\chi}_i = 0 \tag{8}$$

Note that both $\overline{\beta}_i$ and $\overline{\chi}_i$ should be non-negative. Moreover, if either $\overline{\beta}_i = 0$ or $\overline{\chi}_i = 0$, then the case goes back to repetitions of the static partnership game without monitoring. That is, if $\overline{\beta}_i = 0$, then $c'(\hat{e}_i) = \frac{\alpha}{N}\frac{\partial \pi}{\partial e_i}$ which is the effort choice in a static team production as in Holmström (1982). Or if $\overline{\chi}_i = 0$, then no monitoring effort is made since agents have no incentive to monitor without getting compensated for their monitoring effort. With $m_i = 0$ for all i, each agent would make, at most, effort level in the static Nash equilibrium.

However, the repetition of a symmetric static Nash equilibrium is not viable. To see this,

$$\begin{aligned} & c'(\hat{e}_i) &=& \frac{\alpha}{N}\frac{\partial \pi}{\partial e_i} \\ \Leftrightarrow \quad & c'(\hat{e}_i) &=& \frac{\alpha N}{A(N+\alpha)}\frac{\partial}{\partial e_i} f(e_i, ..., e_N) \end{aligned}$$

The RHS is less than 1 by assumption T5, while the LHS is greater than 1 for all $e_i \in [0, 1]$ by assumption of $c'(0) > 1$. Hence, the effort choices in a symmetric Nash equilibrium are $e_i = 0$ for all i and $\pi(N) = -AK$. The payoff to each agent then becomes

$$\frac{\alpha}{N}\pi(N) - c(0) - \frac{A}{N}K = -\frac{\alpha}{N}AK - \frac{A}{N}K < \underline{u}$$

Therefore, the agent is better off not joining the firm.

For the sub-partnership among the agents to be viable, positive monitoring effort should be made. The role of monitoring can be read from (7). To induce higher production effort than that in the static partnership, $\overline{\beta}_i$ should be positive, When $m_i = 0$ for all i, $\overline{\beta}_i$ goes to infinity, which cannot be credibly promised. Therefore, to induce any positive effort, it is necessary to make positive monitoring efforts.

For the bonus to be credible, the sum of maximum bonus payments across the agents should be less than or equal to the total surplus. That is, the following should hold:

$$(RC) \qquad \sum_{i=1}^{N}\{\overline{\beta}_i + \overline{\chi}_i\} \leq \frac{\delta\alpha}{1-\delta}(E(Y_N|\hat{e}) - \sum_{i=1}^{N} c(\hat{e}_i) - \sum_{i=1}^{N}\gamma(\hat{m}_i) - AK - N\underline{u})$$

Then proposition 4 applies and (RC) cannot be satisfied for N large enough. That is, the required amount of bonus increases fast because of the worsening moral hazard problem and as a result, the above (RC) cannot be satisfied. This implies that there exists the finite firm size that maximizes the principal's share of profit.

References

[1] Alchian, A. A., and H. Demsetz (1972), "Production, Information Costs, and Economic Organization," *American Economic Review 62(5)*: 777-95

[2] Arnott, R. and J. E. Stiglitz (1991), "Moral Hazard and Nonmarket Institutions: Dysfunctional Crowding Out or Peer Monitoring?," *American Economic Review 81(1):* 179-90.

[3] Aron, D. J. (1988), "Ability, Moral Hazard, Firm Size, and Diversification," *Rand Journal of Economics 19(1)*: 72-87.

[4] Becker, G. S., and K. M. Murphy (1992), "The Division of Labor, Coordination Costs and Knowledge," *Quarterly Journal of Economics 107(4)*: 1138-1160.

[5] Bull, C. (1987), "The Existence of Self-Enforcing Implicit Contracts," *Quarterly Journal of Economics 102(1)*: 147-60.

[6] Calvo, G. A., and S. Wellisz (1978), "Supervision, Loss of Control, and the Optimum Size of the Firm," *Journal of Political Economy 86(5)*: 942-952.

[7] Camacho, A. and W. D. White (1981),"A Note on Loss of Control, and the Optimum Size of the Firm," *Journal of Political Economy 89(2)*: 407-410.

[8] Che, Y.-K., and S.-W. Yoo (2001), "Optimal Incentives for Teams," *American Economic Review 91(3)*: 525-41.

[9] Clark, J.B. (1908), The Distribution of Wealth: A Theory of Wages, Interest, and Profits. The Macmillan Company.

[10] Coase, R. (1937), "The Nature of the Firm," *Economica 4*: 386-405.

[11] Enright, M. J. (1995), "Organization and Coordination in Geographically Concentrated Industries," in *Coordination and Information*, Ed. N. R. Lamoreaux and D. M. G. Raff, MIT, Cambridge, 1995.

[12] Fama, E. F., and M. L. Jensen (1983), "Separation of Ownership and Control," *Journal of Law and Economics, 26*: 301-325

[13] Fudenberg, D., B. Holmström, and P. Milgrom (1990), "Short-Term Contracts and Long-Term Agency Relationships," *Journal of Economic Theory, 51*: 1-31

[14] Fudenberg, D., D. Levine, and E. Maskin (1994), "The Folk Theorem with Imperfect Public Infor-mation," *Econometrica 62(5)*: 997-1039.

[15] Grossman, S., and O. Hart (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy 94(4)*: 691-719.

[16] Hansmann, H. (1996), The Ownership of Enterprise, Belknap-Harvard.

[17] Hart, O., and J. Moore (1990), "Property Rights and the Nature of the Firm," *Journal of Political Economy 98(6)*: 1119-58.

[18] Holmström, B. (1979), "Moral Hazard and Observability," *Bell Journal of Economics 10(1)*: 74-91.

[19] Holmström, B. (1982), "Moral Hazard in Teams," *Bell Journal of Economics 13*:324-340.

[20] Holmström, B., and P. Milgrom (1987), "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica 55(2)*: 303-328.

[21] Holmström, B., and J. Roberts (1998), "The Boundaries of the Firm Revisited," *Journal of Economic Perspectives 12(4)* : 73-94.

[22] Itoh, H. (1992), "Cooperation in Hierarchical Organizations: An Incentive Perspective," *Journal of Law, Economics and Organization 8(2)*: 321-345

[23] Jewitt, I. (1988), "Justifying the First-Order Approach to Principal-Agent Problems," *Econometrica 56(5)*: 1177-1190.

[24] Kaldor, N. (1934), "The Equilibrium of the Firm," *Economic Journal 44*: 60-76.

[25] Kandel, E., and E. Lazear (1992), "Peer Pressure and Partnership," *Journal of Political Economy 100(4)*: 1119-58.

[26] Keren, M., and D. Levhari, (1983), "The Internal Organization of the Firm and the Shape of Average Costs," *Bell Journal of Economics 14*: 474-488.

[27] Knez, M., and D. Simester (2001), "Firm-Wide Incentives and Mutual Monitoring at Continental Airlies," *Journal of Labor Economics 19(4)*: 743-772.

[28] Levin, J. (2002a), "Multilateral Contracting and Employment Relationship," *Quarterly Journal of Economics 117*: 1075-1103.

[29] Levin, J. (2002b), "Relational Incentive Contracts," *American Economic Review, Forthcoming.*

[30] Levine, D. I., and L. D. Tyson (1990), "Participation, Productivity, and the Firm's Environment," in *Paying for Productivity: A Look at the Evidnece*, Ed. A. Blinder, The Brookings Institution, Washington, D.C., 1990.

[31] Lucas, R. (1978), "On the Size Distribution of Business Firms," *Bell Journal of Economics 9*: 508-523.

[32] Ma, C. (1988), "Unique Implementation of Incentive Contracts with Many Agents," *Reveiw of Economic Studies 55(4)*: 555-571.

[33] MacLeod, W.B., and J. Malcomson (1988), "Reputation and Hierarchy in Dynamic Models of Employment," *Journal of Political Economy 96(4)*: 832-54.

[34] MacLeod, W.B., and J. Malcomson (1989), "Implicit Contracts, Incentive Compatibility, and Invol-untary Unemployment," *Econometrica 57(2)*: 447-80.

[35] MacLeod, W.B., and J. Malcomson (2000), "Motivation and Markets," *American Economic Review 88(3)*: 388-411.

[36] Matsui, A., and A. Postlewaite (2000), "Specialization of Labor and Distribution of Income," *Games and Economic Behavior 33*: 72-89.

[37] McAfee, R.P., and J. McMillan (1995), "Organizational Diseconomies of Scale," *Journal of Economics and Management Strategy 4(3)*: 399-426

[38] Mookherjee, D. (1984), "Optimal Incentive Schemes with Many Agents," *Review of Economic Studies 51(3)*: 433-46.

[39] Nalbantian, H. R. (1988), "Incentive Compensation in Perspective," in *Incentives, Cooperation, and Risk Sharing*, Ed. H. R. Nalbantian, Rowman and Littlefield, Savage, 1988.

[40] Pearce, D.G., and E. Stacchetti (1998), "The Interaction of Implicit and Explicit Contracts in Repeated Agency," *Games and Economic Behavior 23(1)*: 75-96.

[41] Rajan, R. G. and L. Zingales (1998), "Power in a Theory of the Firm," *Quarterly Journal of Economics 113(2)*: 147-60.

[42] Rajan, R. G. and L. Zingales (2001), "The Firm as a Dedicated Hierarchy: A Theory of the Origins and Growth of Firms," *Quarterly Journal of Economics 116(1)*: 147-60.

[43] Rasmusen, E., and Zenger (1990), "Decresing Returns to Scale in Employment Contracs," *Journal of Law, Economics and Organization 6(1)*: 65-92.

[44] Rayo, L. (2002), "Relational Team Incentives and Ownership," mimeo.

[45] Rosen, S. (1979), "Substitution and Division of Labour," *Economica 45(179)*: 235-50.

[46] Rosen, S. (1983), "Specialization and Human Capital," *Journal of Labor Economics 1(1)*: 43-49.

[47] Williamson, O. E. (1967), "Hierarchical Control and Optimum Firm Size," *Journal of Political Economy 75(2)*: 123-38.

[48] Williamson, O. E. (1985), *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: The Free Press, 1975.

[49] Ziv, A. (1993), "Performance Measures and Optimal Organization," *Journal of Law, Economics and Organization 9(1)*: 30-50.