# Incentives, Supervision, and Limits to Firm Size

Ayush Gupta*
*Boston University*

Last Updated: 8 February 2024

**Abstract**

This paper studies how incentive problems affect the size and structure of a firm (or any other organization) with one principal and many agents. We develop a model in which worker effort, supervision, and wages are determined endogenously. Our model generalizes much of the existing literature by making minimal assumptions about the production technology and the nature of supervision. Using a novel optimization technique, we establish necessary and sufficient conditions for the size of a firm to not be limited by incentive problems. We show that firms are limited to a positive, but finite, number of workers under reasonable assumptions. Firm size is unbounded only when worker productivity is sufficiently greater than the disutility of effort.

*I am indebted to Andy Newman and Dilip Mookherjee for their advice and support throughout this project. I would also like to thank seminar participants at Boston University, Boston College, IOEA 2024, and the Midwest Economic Theory Conference for helpful comments and suggestions.

# 1 Introduction

Economic models typically treat firms as singular agents trying to maximize profits. But firms are not monolithic entities; they are made up of individuals with diverse objectives. This creates an incentive problem when an individual's objective conflicts with the firm's. Established results tell us that incentive problems lead to individuals taking actions which are not optimal from the firm's perspective.[1] This is often called "loss of control" in the literature. An open question in organization economics is how this loss of control changes as the size of a firm increases. If it is increasing, then incentive problems alone may impose a limit on the size of a firm.[2]

Our object of study is any contractual arrangement with one principal and many agents. While we present our results in the context of a firm, our model describes many other types of organizations. Conversely, not every organization which meets the legal definition of a firm will satisfy our assumptions. We start by showing that it is not profitable for the owner to directly supervise an infinite number of workers. This forces firms to adopt a hierarchical structure - the owner supervises a subset of workers who, in turn, supervise other workers and so on. We then establish necessary and sufficient conditions for a firm to continue this process *ad infinitum*. These conditions show that, under reasonable assumptions, incentive problems are increasing in the size of a firm. When that happens, firms are limited to a positive but finite number of workers.

Unlike much of the existing literature, we find that incentive problems alone can limit the size of a firm. This literature dates back to at least Calvo and Wellisz (1978) (hereafter "CW78") which endogenized loss of control using a model of supervision and wage incentives. They demonstrated that, in general, incentive problems do not limit the size of a firm. A firm can grow infinitely large, as long as it is profitable to hire a single worker. In the CW78 model, firm size is limited only if workers can identify the times at which their performance is being monitored. Subsequent papers largely supported the CW78 result, showing that incentive problems do not limit firm size, except under very specific assumptions. For example, Qian (1994) assumed a special production function such that the profit generated by a worker is a multiplicative function of his own effort and the effort exerted by all of his (direct and indirect) supervisors. Unless all workers exert full effort, this assumption limits firm size by causing productivity to decline as the firm adds more hierarchical layers. Similarly, Keren and Levhari (1983) showed that the optimal firm size is finite but assumed that communication within a firm is costly. On the other hand, both Datta (1996) and Tsumagari (1999) showed that firm size remains unbounded even when supervisory effort is not directly observable.

[1] See Mirrlees (1976); Shavell (1979); Holmström (1979); Grossman and Hart (1983).

[2] See Williamson (1967) for a formal treatment of how increasing loss of control limits the size of a firm.

Our necessary condition shows that the growth of a firm is not constrained by incentive problems only if it is profitable to re-organize the firm as a stationary hierarchy - a hierarchy in which every worker's action is the same.[3] That is only possible when the productivity of effort is "sufficiently" greater than the disutility of effort.[4] Intuitively, this is because incentive problems impose an agency cost on the firm. So it is not sufficient for a worker's production to be greater than his wage (which is proportional to the disutility of effort). Worker productivity needs to be high enough to offset the agency cost.

The proof of our necessary condition constitutes a significant contribution in its own right. We are able to set up the firm's optimization problem as an infinitely repeated dynamic programming problem. There is no discounting so this problem cannot be solved using standard techniques. Instead, we apply the Kuhn-Tucker theorem in a novel way to show that the profit generated by a stationary hierarchy is, at most, finitely less than the profit generated by any arbitrary hierarchy. This approach can be applied to a large variety of problems where the objective is to obtain a bound on the value function.

Our model is built on two key assumptions. First, we assume that effort is costly so workers shirk on the job unless supervised by their manager. Second, we assume that supervision requires time which might otherwise be spent on profit generating activities. This results in a non-trivial trade-off because each individual has a finite allocation of time. Beyond these two assumptions, we impose minimal restrictions on the production technology, the nature of supervision, or the type of employment contracts. Consequently, our model generalizes much of the existing literature and our results are widely applicable.

This paper also contributes to a much larger literature initiated by Coase (1937). Coase proposed that firms exist to avoid transaction costs but cannot grow arbitrarily large because of decreasing returns to the entrepreneur function. Since then, both the costs and the benefits of a firm have been extensively studied. This paper focuses on the former. To that end, we take the benefits of a firm as given and study one of the factors which might limit its size. We note that our model does not speak to the firm versus market distinction. Our results can help determine the optimal size of a firm, but not its scope.

One branch of the literature focuses on incomplete contracts. Grossman and Hart (1986) and Hart and Moore (1990) showed that incomplete contracts create hold-up problems which limit firm size because ownership needs to be aligned with the incentive to invest. Our paper does not assume incomplete contracts. Instead, we assume that even complete contracts are costly to enforce in the

[3] We do not claim that stationary hierarchies are optimal, nor do we restrict our analysis to such hierarchies.

[4] In comparison, if there are no incentive problems, firm size is unbounded whenever productivity is even a little greater than the disutility of effort.

presence of incentive problems. In the limit case, the cost can be interpreted as the time needed to verify that the counter-party has fulfilled their obligations under the contract. This cost can be very small but we argue that it will never be zero. For example, if an employee is paid a piece rate, the owner will need to verify the quantity and quality of the goods produced. Doing so might be very simple but it will always require some positive amount of time.

A different branch of the literature assumes that communication within a firm is costly. Radner (1993), Bolton and Dewatripont (1994), Garicano (2000), and many others have studied how costly communication affects the internal organization of a firm and shown how it may limit firm size. We consider these works to be complementary to our own. We assume that there are no communication costs only because we want to isolate the effects of incentive problems. Our results show that, even as communication costs approach zero, the optimal size and structure of a firm may be constrained by incentive problems alone.

Outside the firm setting, our assumptions require only that enforcing a contract costs time, of which the principal has a finite allocation. Then our results show that a principal cannot contract with an infinite number of agents. She will eventually need to allow for sub-contractors, sub-sub-contractors and so on. The number of agents who can be contracted with in this manner is infinite only if our necessary condition is satisfied. Otherwise, incentive problems limit the size of the contractual arrangement.

For ease of exposition we start by presenting our results in the CW78 framework. Section 3 uses a simple example to show how, even in this simplified setting, incentive problems can limit firm size. Section 4 describes our general model and states our main results. Section 5 concludes.

## 2 Simplified Model

This section describes the CW78 model to provide intuition. Our general model relaxes many of the assumptions made here and can be applied to a much larger class of problems.

The object of study is a hierarchical contractual arrangement with one principal and many agents. To remain consistent with the existing literature, we will call this arrangement a firm. However, we stress that our assumptions are not equivalent to the legal definition of a firm. Consequently, our model describes many non-firm organizations but not every firm satisfies our assumptions.

We will assume that the solitary principal (hereafter "owner") wants to maximize profit. She can operate alone or hire from an infinite pool of identical agents (hereafter "workers"). The profit

generated by each worker is given by $f(p)$ where $p$ is the worker's level of productive effort. We assume:

1. $f(p)$ continuous, increasing, and bounded
2. $f(0) = 0$

Notice that the profit generated by a worker depends only on his own level of productive effort. That means that there is no complementarity in worker effort. Furthermore, the total profit generated by the firm is the sum of the profit generated by each worker and the owner.[5] These assumptions ensure that, in the absence of incentive problems, the firm exhibits constant returns to scale.[6]

The von Neumann-Morgenstern utility index of each worker is given by $u(w, e)$ where $w \geq 0$ is the wage received by the worker and $e \in [0, 1]$ is the level of total effort exerted by him. Here, effort can be interpreted as the fraction of the day spent working rather than shirking. We assume:

1. $u(e, w)$ continuous
2. $u(w, e)$ increasing in $w$
3. $u(w, e)$ decreasing in $e$
4. $u(0, e) \leq 0$ for all $e$
5. Workers have a positive outside option which gives them $\underline{u} > 0$ utility

If $\underline{u} = 0$ then it is possible for the firm to hire an infinite number of workers who exert zero effort and receive zero wage. Assuming $\underline{u} > 0$ avoids this trivial case. The reader should interpret $\underline{u}$ as being arbitrarily close to 0.

It is obvious that a worker who is offered a fixed wage[7] will maximize her utility by exerting zero effort. The owner can resolve this incentive problem by offering employment contracts which are contingent on effort. But such a contract will not be effective unless the owner observes effort. Observing effort requires supervision which is both costly and imperfect. More concretely, suppose that $s \in [0, 1]$ is the level of supervision received by a worker. Then $s$ is the probability that the

[5] These assumptions are consistent with the "putting out" system of production.

[6] We are being a little imprecise here. The firm will only exhibit constant returns to scale when the owner's production is zero. When the owner's production is positive, total production will increase linearly in the number of workers.

[7] A wage is said to be fixed if it does not depend on the worker's level of effort.

worker's effort is perfectly observed and also the fraction of the day that the owner spends on supervising said worker. Intuitively, the owner knows more about a worker's effort level when she spends more time supervising that worker. The general model in Section 4 relaxes this assumption and allows for other forms of supervision.[8]

The trade-off faced by the owner is that time spent on supervising a worker cannot be spent on generating profit. So one can think of supervision as being unproductive to the extent that it does not (directly) generate profit. However, the owner can instruct a worker to supervise other workers, allowing for hierarchical firms with managers.

In keeping with the CW78 model, we assume that the owner offers a wage $\overline{w}$. The worker receives $\overline{w}$ if his effort level is not observed. If the worker's effort is observed, he receives $e\overline{w}$ where $e$ is his level of effort.[9]

If his participation constraint is satisfied, the worker picks $e$ to maximize his expected utility:

$$U = \max_{e} su(e\overline{w}, e) + (1 - s)u(\overline{w}, e)$$

Let $\mathbf{E}$ be the set of maximizers. In general, $\mathbf{E}$ will not be a singleton but it will be a compact set. For the remainder of this paper, we will assume that the worker's chosen level of effort is $e^* = \max(\mathbf{E})$.[10] This assumption allows us to define a function $e(\overline{w}, s)$ which gives the worker's choice of effort, conditional on $\overline{w}$ and $s$. Use $e(\overline{w}, s) = \phi$ to denote the case when the worker's participation constraint is not satisfied and he chooses his outside option.

Knowing $e(\overline{w}, s)$, the owner can induce any feasible level of effort by offering a wage $\overline{w}(e, s)$ where $e$ is the desired effort and $s$ is the level of supervision received by the worker. Given some $\overline{w}(e, s)$, the expected wage that the owner will have to pay is:

$$w(e, s) = se\overline{w}(e, s) + (1 - s)\overline{w}(e, s)$$

We call $w(e, s)$ the expected wage function. Notice that it will have the following properties:

- Increasing in $e$ because workers need to be compensated for the disutility of effort
- $w(e, s) \to \infty$ as $s \to 0$ because unsupervised workers receive a fixed wage and never exert any

[8] Many existing papers (including CW78) justify effort contingent contracts on the grounds that it is often easier for the owner to verify effort rather than output. We allow for output contingent contracts in Section 4 but note here that effort contingent contracts typically make it easier to achieve the first-best outcome (relative to output contingent contracts).

[9] This contract is not optimal from the owner's perspective because it is possible to induce a higher level of effort for the same expected wage. Our general model relaxes this assumption and allows for arbitrary contracts.

[10] We consider this to be a conservative assumption. If we allow the agent to pick a lower level of effort, it becomes even more likely that incentive problems will limit firm size.

effort

- $w(e, s) \geq \underline{w} > 0$ for all $e$ because workers have an outside option which gives them $\underline{u} > 0$ utility

A worker can split his effort between activities which generate profit (productive effort) and supervising other workers (supervisory effort). We assume that effort is infinitely divisible so that $e = p + s$ where $p$ is the level of productive effort and $s$ is the level of supervisory effort. We also assume that, conditional on the level of effort, workers are indifferent between productive work and supervisory work. It is easily verified that this indifference assumption is unnecessary for our results.

In this model all organizational decisions are made by the owner. The owner decides the size and structure of the firm i.e. how many workers to hire and who supervises whom. The owner also decides how much effort to induce from each worker and how to split that effort between productive work and supervisory work. The managers serve only to verify the level of effort exerted by their subordinates. In particular, we are assuming that managers cannot mis-report what they observe so there is no possibility of collusion between workers and managers.

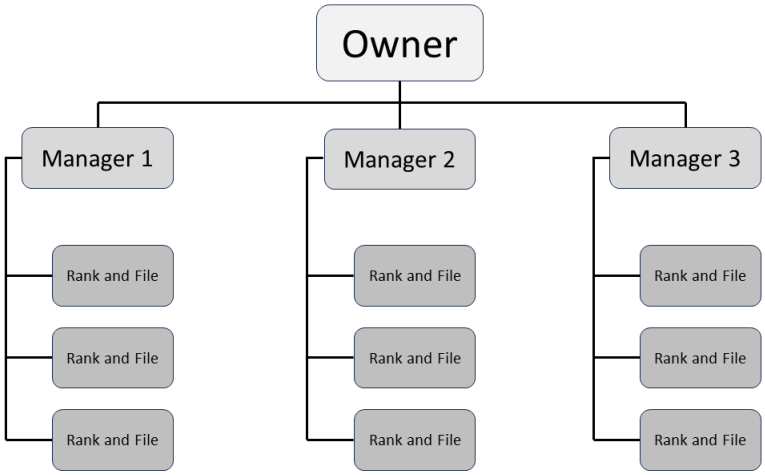


Figure 1: Example Firm Hierarchy

Before moving on to the results, it is helpful to introduce an example and some definitions. Figure 1 shows the organizational structure of a hypothetical firm with twelve employees. In this firm, the

owner directly supervises three employees (labelled "Managers" in Figure 1). In turn, each manager supervises three employees (labelled "Rank and File" in Figure 1). The rank and file workers do not supervise anyone so all of their effort is necessarily productive effort. However, the owner and the three managers may split their effort between productive and supervisory work in any ratio.

Note that we use the words worker and employee interchangeably to refer to any agent who is not the owner of the firm. The words manager and supervisor are used interchangeably to refer to any agent who supervises another agent. So all managers are workers but the converse is not true.

Layer $n$ is defined as the set of workers who are $n$ degrees of separation from the owner. The owner is always Layer 0; Layer 1 is the set of workers who are directly supervised by the owner (the three managers in Figure 1); Layer 2 is the set of workers whose supervisor is directly supervised by the owner (the nine rank and file workers in Figure 1); and so on.

A branch is defined as a (possibly infinite) sequence of workers such that worker $i$ is supervised by worker $i-1$ and worker 1 is supervised by the owner.

A firm is said to grow horizontally when the number of layers remains constant but the number of workers increases i.e. workers are added to some (or all) layers. A firm is said to grow vertically when the number of layers increases.

For simplicity we will assume that the owner always exerts $e = 1$ effort.[11] She will generate $f(1)$ units of profit if she operates alone. If the owner hires exactly one worker, the firm will generate:

$$f(1-s) + f(e) - w(e,s)$$

Where the owner generates $f(1-s)$ and the worker generates $f(e)$.[12] The owner will hire a worker if and only if there exists some $e \in [0,1]$ and $s \in [0,1]$ such that this profit is greater than the profit generated by the owner operating alone. That gives:

$$\begin{aligned} f(1) &< f(1-s) + f(e) - w(e,s) \\ f(1) - f(1-s) &< f(e) - w(e,s) \end{aligned}$$

This observation provides some useful intuition. $f(e) - w(e,s)$ is the net profit generated by the worker after accounting for his wage. $f(1)$ is the profit generated by the owner when she operates alone and $f(1-s)$ is the profit she generates when she has to supervise the worker. So $f(1)-f(1-s)$ is the reduction in the owner's productivity and can be interpreted as the opportunity cost of

[11] This assumption is only made for ease of exposition. None of our results depend on the level of effort exerted by the owner.

[12] There is no one for the worker to supervise so all of his effort is productive effort.

hiring and supervising a worker. This cost is a result of the incentive problem which necessitates supervision.

Because of the incentive problem, it is not enough for the net profit generated by a worker to be positive. The owner will not hire a worker unless he generates enough profit to make up for the opportunity cost of supervising him. As a result, the first best outcome is not possible in this model: there will be some situations where the net profit generated by a worker is positive but he is not hired because the cost of supervising him is too high.

Proving the main theorems requires three lemmas which are stated below. All proofs are included in the appendix.

**Lemma 1.** *The span of control is finite i.e. one manager cannot (profitably) supervise an infinite number of workers.*

The proof of Lemma 1 leverages the fact that no manager can exert an infinite amount of supervisory effort. As a manager's number of (direct) subordinates increases, the level of supervision received by individual subordinates must decrease. Since $w(e, s) \to \infty$ as $s \to 0$, the expected wage required to induce any positive level of effort will eventually be so large that the net profit generated by the subordinate will be negative. This will allow the firm to increase total profit by decreasing the number of workers.

**Lemma 2.** *Any firm with a finite number of layers will have a finite number of workers.*

The proof of Lemma 2 uses Lemma 1 to show that every layer of a firm must be finitely large. In other words, incentive constraints limit the ability of a firm to grow horizontally. More importantly, the contra-positive of Lemma 2 states that an infinitely large firm must have an infinite number of layers i.e. a firm must grow vertically if it is to expand beyond finite size.

**Lemma 3.** *A firm with $N$ layers will have at least one branch with $N$ workers.*

The significance of Lemma 3 is that it can be combined with Lemma 2 to show that any infinitely large firm must have at least one infinitely long branch. This is critical for the proof of the main

theorems which are stated below.

**Theorem 1.** *A firm can grow infinitely large and generate infinite profit if there exists some* $e^* \in [0,1]$ *and* $s^* \in [0, e^*]$ *such that:*

$$f(e^* - s_1^*) > w(e^*, s_2^*)$$

*AND*

$$s_1^* = s_2^* = s^*$$

$e^*$ is the total effort exerted by a worker; $s_2^*$ is the supervision received by him; $s_1^*$ is the level of effort the worker spends on supervising his own subordinate; and $e^* - s_1^*$ is the worker's level of productive effort.

Theorem 1 proves a sufficient condition for incentive constraints to not limit firm size. Theorem 2 proves that a slightly weaker condition is necessary.

**Theorem 2.** *A firm can grow infinitely large and generate infinite profit only if there exists some* $e^* \in [0,1]$ *and* $s^* \in [0, e^*]$ *such that:*

$$f(e^* - s_1^*) \geq w(e^*, s_2^*)$$

*AND*

$$s_1^* = s_2^* = s^*$$

Superficially, the necessary and sufficient conditions require that the profit generated by a worker (given by $f(e^* - s_1^*)$) is greater than his wage (given by $w(e^*, s_2^*)$). The meaningful insight comes from the constraint $s_1^* = s_2^*$. This constraint represents the fact that it is not enough for a worker's production to be greater than his wage. Incentive problems will limit firm size unless the net profit generated by a worker is positive even when his level of supervisory effort is the same as the level of supervision received by said worker.[13]

The intuition is that incentive problems necessitate supervision, which means that the total cost of hiring a worker is greater than his wage. This limits firm size unless workers can compensate the firm for the additional cost of supervision. The necessary condition serves to specify the requisite level of compensation. One interpretation is that workers need to generate positive profit while supplying as much supervision as they receive. Then we can say that the net supervision received by the workers is zero so net profit generated by them is positive even after accounting for the cost of supervision.

[13]Recall that $f$ is an increasing function so the necessary and sufficient conditions will be satisfied if there exists some $e^*$ and $s_1^* > s_2^*$ such that $f(e^* - s_1^*) > w(e^*, s_2^*)$.

An alternate interpretation comes from rearranging the necessary condition as follows:

$$f(e^*) - f(e^* - s^*) \leq f(e^*) - w(e^*, s^*)$$

Notice that the right hand side of the inequality represents the net profit generated by a worker who receives $s^*$ supervision but does not supervise anyone else. The left hand side represents the decrease in this worker's production if he himself were to exert $s^*$ supervisory effort. It can be interpreted as the "value" of the supervision received by the worker.[14] Then the necessary condition can be interpreted as saying that incentive problems limit firm size unless the net profit generated by a worker is greater than the value of the supervision received by him.

The proof of Theorem 1 utilizes a replication argument. Suppose the sufficient condition holds. Then a worker can generate positive profit while receiving $s^*$ supervision and exerting $s^*$ effort on supervising a subordinate. But then the subordinate can replicate his supervisor: he is receiving $s^*$ supervision so he can also generate positive profit while exerting $s^*$ effort on supervising his own subordinate. This replication process can be continued *ad infinitum* to generate an infinitely large firm in which every employee is "doing the same thing" and generating positive net profit.

The proof of Theorem 2 is rather more involved. The intuition can be understood by focusing on what happens when the inequality is only satisfied for $s_1^* < s_2^*$ i.e. a worker can generate positive profit only when his subordinate receives strictly less supervision than what the worker receives. But then the subordinate cannot replicate his manager. In particular, he will either exert less effort or receive a higher wage. This will be true at all layers of the hierarchy. Thus, we will eventually reach a layer where the workers are generating negative profit for the firm.[15]

Note that Theorem 2 does not claim that a stationary hierarchy is optimal, nor are we restricting our analysis to firms which are organized as such. Theorem 2 only states that, if incentive problems do not limit firm size, then the firm can be reorganized as a stationary hierarchy and remain profitable. In practice, a firm might find it more profitable to employ specialist managers i.e. employees who spend all of their effort on supervising other workers. This possibility is not excluded from our analysis.

Theorems 1 and 2 establish conditions under which incentive problems do not limit firm size. Firm size is not limited when the condition in Theorem 1 is satisfied. Conversely, if firm size is not limited by incentive problems, then the condition in Theorem 2 must be satisfied. Consequently, Theorems 1 and 2 provide a useful benchmark for testing if incentive problems constrain the growth of a firm.

[14] Under this definition the value of supervision is different from its opportunity cost.

[15] We cannot rule out the possibility that the profit generated by each worker approaches zero but never becomes negative. That is why the necessary condition is stated in terms of a weak inequality while the sufficient condition is stated in terms of a strong inequality.

They can be used in large variety of situations where economists are potentially concerned about incentive problems causing dis-economies of scale.

# 3 Example

The significance of our results can be illustrated with a simple example. Assume that:

$$f(p) = p^{\frac{1}{3}}$$
$$u(e, w) = w - e^2$$
$$\underline{u} = 0.1 \Rightarrow \underline{w} = 0.1$$

Assume that employment contracts follow the CW78 model i.e. workers are paid $\overline{w}$ if their effort is not observed and they are paid $e\overline{w}$ if their effort is observed. When their participation constraint is satisfied, workers will exert $e = \frac{1}{2}s\overline{w}$ and the expected wage function will be given by:

$$w(e, s) = \frac{2e}{s} - 2e(1 - e)$$

When the participation constraint is not satisfied, the owner will have to offer a higher $\overline{w}$. Therefore, the given $w(e, s)$ is a lower-bound and not the true expected wage function. However, this lower-bound is sufficient for our example so, for ease of exposition, we will not use the true expected wage function.

The owner will hire at least one worker if there exists some $e$ and $s$ such that:

$$(1 - s)^{\frac{1}{3}} + e^{\frac{1}{3}} - \frac{2e}{s} + 2e(1 - e) > 1$$

It is easily verified that there are many $(e, s)$ which satisfy this condition. One example is $e = 0.1$ and $s = 0.5$.

Therefore, it is profitable for the owner to hire at least one worker. However, we claim that the necessary condition from Theorem 2 is not satisfied so this firm cannot grow infinitely large.

*Proof.* Fix any $e \in (0, 1]$. Theorem 2 requires that there exists some $s \in [0, e]$ such that:

$$(e - s)^{\frac{1}{3}} \geq \frac{2e}{s} - 2e(1 - e)$$

Notice that $\frac{2e}{s} - 2e(1 - e) \geq \frac{3}{2}$ when $s = e$ and $\frac{2e}{s} - 2e(1 - e) = \infty$ when $s = 0$.

We can see that $\frac{2e}{s} - 2e(1 - e)$ is a continuous function which is always decreasing in $s$ (for any

fixed $e$).

Then we know that $\frac{2e}{s} - 2e(1-e) \in [\frac{3}{2}, \infty)$.

But we also know that $(e-s)^{\frac{1}{3}} \in [0,1]$ so the condition in Theorem 2 is never satisfied. □

Recall that Theorem 2 provides a necessary condition for firm size to *not* be limited by incentive problems. We have thus shown that our hypothetical firm *will* be limited by incentive problems.

In our example, it is profitable for the firm to hire at least one worker but incentive problems prevent it from growing infinitely large. Therefore, a finitely large firm will be optimal in this example, which is never the case in the CW78 model.[16]

Notice that our example is constructed to fit the CW78 model in all respects, except for one assumption: we allow $f$ to be non-linear. We believe that this is a reasonable relaxation because assuming a linear $f$ is not appropriate in many situations. For example, any industry in which worker fatigue is a concern is likely to exhibit a concave $f$ (which is what we assume here).

Intuitively, the CW78 result fails because it is not profitable for Worker 1 to replicate the owner. More precisely, it is profitable for the owner to exert $s$ on supervising a worker who exerts effort $e$; but it is not profitable for Worker 1 to exert $s$ on supervising a worker who exerts effort $e$. This is because Worker 1 is not productive enough to generate positive profit while supplying as much supervision as he receives.

It is important to note that our choice of $f$ does not automatically limit the size of the firm. Even though the individual production function is assumed to be concave, the firm as a whole need not exhibit decreasing returns to scale. We illustrate this point by considering the extreme case when there is no incentive problem and supervision is unnecessary.

Assume that workers are paid just enough to be indifferent between working and their outside option so:

$$w(e) = e^2 + 0.1$$

Then the owner will hire a worker as long as there exists some $e$ such that:

$$e^{\frac{1}{3}} > e^2 + 0.1$$

It is easily verified that many such $e$ exist (one example is $e = 0.3$). So the owner will hire at least

[16] It is possible to show that a two layer firm will be optimal in this example. The optimal firm will have the owner directly supervising a positive but finite number of workers.

one worker and the resulting firm will generate more profit than a firm with no workers. But now that supervision is unnecessary, there is nothing preventing the owner from hiring more workers. In particular, every additional worker will add a positive amount to the total profit generated by the firm. The owner will want to hire an infinite number of workers which will yield infinite profit for the firm. Therefore, in the absence of incentive problems, the optimal size of the firm is unbounded. It is the necessity of supervision which imposes a binding constraint on firm size, not our choice of $f$.

Note that we assume individual worker output is aggregated linearly and that there is no complementarity in worker effort precisely so that the firm exhibits constant returns to scale in the absence of incentive problems.[17] Many existing papers in the literature relax one (or both) of these assumptions by adding coordination costs or assuming that a worker's output depends on the actions of multiple workers. But then it is not possible to make the argument that incentive problems are limiting firm size. It could very well be the case that firm size is limited even in the absence of incentive problems.

# 4 General Model

This section presents our general model which encompasses the Calvo-Wellisz setting and the simple model presented in Section 2. We start by relaxing two significant assumptions which we have maintained so far:

1. We now allow the principal to offer any arbitrary contract which satisfies the limited liability constraint - meaning that the agent's wage cannot be negative. In the firm setting, this means that the worker can be dismissed but he cannot be compelled to pay the owner.

2. We no longer assume that the agent's effort is perfectly observed with probability $s$. Instead, we assume that, given a level of effort $e$, the principal observes a signal $y \in Y(e) \subseteq Y$ where the conditional probability of observing $y$ given $e$ is a function of the level of supervision $s$.[18]

   We continue to assume that some supervision is necessary to learn anything about the agent's level of effort. Formally, we assume that, as $s \to 0$, for every $y \in Y$ and every $e \neq e'$, we get $\Pr[y|e] \to \Pr[y|e']$.

[17] The caveat from Footnote 6 still applies.

[18] We assume that the set $Y(e)$ always includes $\phi$ and $e$. Then it is easy to see that the monitoring technology in Sections 2 and 3 is allowed under our assumptions.

This framework allows us to extend our model to a large set of contracts and monitoring technologies. Of particular interest is a monitoring technology which allows the principal to perfectly observe the agent's effort when the level of supervision exceeds some minimum threshold ($s \geq \underline{s} < 1$). Such technologies represent a situation in which the principal need only spend a small amount of time to verify that the agent has fulfilled his contractual obligations. This assumption is consistent with many important settings and, as long as $\underline{s} > 0$, our results will continue to hold.

In general, the agent's wage will be a random function of his effort and the level of supervision. This function will be jointly determined by the employment contract and the monitoring technology. We define $\psi : (e, s) \to \Delta w$ to be the mapping from effort and supervision to distributions over wages. Then the agent will pick $e$ to maximize his expected utility:

$$U = \max_{e} \mathbb{E}_{w \sim \psi}[u(e, w)]$$

As before, if there is more than one solution, we assume that the agent picks the highest effort level which maximizes his utility.[19] This allows us to define the expected wage function $w(e, s)$ like we did in Section 2. Note that it will continue to have the following properties:

- Agent utility is decreasing in effort. So $w(e, s)$ is increasing in $e$ because agents need to be compensated for the disutility of effort.
- Some positive level of supervision is necessary to incentivize an agent. So $w(e, s) \to \infty$ as $s \to 0$.
- Agents have a positive outside option which gives them $\underline{u} > 0$ utility. So $w(e, s) \geq \underline{w} > 0$ for all $e$.[20]

Before we proceed, we want to emphasize that our results are not restricted to settings with effort contingent contracts. In particular, our model can easily accommodate various types of output contingent contracts. That is because any contract must ultimately induce a mapping between an agent's wage and his level of effort. If the two are uncorrelated, the agent will maximize his utility by exerting zero effort. Because we impose minimal assumptions on $\psi$, it can be adjusted to capture the relationship between effort and wage under a variety of different contracts.

For example, consider a contract under which the agent's wage is determined by his subordinates' output. Then the agent's wage can be expressed as $\psi(f(e'(e)), s)$ where $e'(e)$ is the mapping from

[19] An additional assumption is required to ensure that at least one solution always exists. The assumption and the proof of existence are detailed in the appendix.

[20] As before we only assume $\underline{u} > 0$ to avoid an infinite firm in which all agents exert zero effort and receive zero wage. The reader should interpret $\underline{u}$ as being arbitrarily close to 0.

the agent's effort $e$ to his subordinates' level of effort $e'$. This type of contract will not change any of our results, as long as our other assumptions are satisfied.

To summarize, there are two key assumptions underpinning our model. The first is that effort is costly so a worker's wage must depend on his level of effort. This may be directly (in case of effort contingent contracts) or indirectly (in case of output contingent contracts). The second key assumption is that enforcing employment contracts requires some supervision, which is costly. The model remains agnostic about the specific relationship between wage and effort, as well as the extent of supervision needed.[21]

With these assumptions in place, we are able to prove all of our results as presented in Section 2. All proofs are included in the appendix but we restate our results for the reader's convenience.

**Lemma 1.** *The span of control is finite i.e. one manager cannot (profitably) supervise an infinite number of workers.*

**Lemma 2.** *Any firm with a finite number of layers will have a finite number of workers.*

**Lemma 3.** *A firm with $N$ layers will have at least one branch with $N$ workers.*

**Theorem 1.** *A firm can grow infinitely large and generate infinite profit if there exists some $e^* \in [0,1]$ and $s^* \in [0, e^*]$ such that:*

$$f(e^* - s_1^*) > w(e^*, s_2^*)$$

*AND*

$$s_1^* = s_2^* = s^*$$

**Theorem 2.** *A firm can grow infinitely large and generate infinite profit only if there exists some $e^* \in [0,1]$ and $s^* \in [0, e^*]$ such that:*

$$f(e^* - s_1^*) \geq w(e^*, s_2^*)$$

*AND*

$$s_1^* = s_2^* = s^*$$

We note that our main results are stated in terms of the expected wage function which is endogenously determined. The benefit of doing so is that it allows us to provide a general result which is independent of the choice of employment contract, the utility function, and the monitoring technology. To do otherwise will result in a different condition for every possible combination of the three. This is both infeasible and uninformative.

[21] Note that these assumptions are equivalent to the ones stated in the introduction.

Moreover, the utility function and the monitoring technology are both exogenous, so there exists a simple relationship between the expected wage function and the primitives. That is because the employment contract assumed by Qian (1994) is optimal from the principal's perspective. Under this contract, the principal specifies a minimum acceptable level of effort $\underline{e}$. The agent receives wage $\overline{w}$ if he exerts $e \geq \underline{e}$. If the agent is observed to exert $e < \underline{e}$, his wage is zero. Then the agent exerts effort $\underline{e}$ if and only if:

$$u(\overline{w}, \underline{e}) \geq (1 - \pi(s))u(\overline{w}, 0)$$

and

$$u(\overline{w}, \underline{e}) \geq \underline{u}$$

Where $\pi(s)$ is the probability that the agent receives zero wage when he exerts no effort. This is a function of the level of supervision received by the agent and the monitoring technology used by the principal. Then, given some level of supervision $s$ and a desired level of effort $e$, the optimal wage $w^*$ solves:

$$u(w^*, 0) = \frac{u(w^*, e) - u(w^*, 0)}{\pi(s)}$$

if $u(w^*, e) > \underline{u}$ and otherwise

$$u(w^*, 0) = u(w^*, e) - u(w^*, 0) + \underline{u}$$

In equilibrium, the agent will always exert effort $e$ and receive wage $w^*$. The expected wage function is then given by $w(e, s) = w^*$. This allows us to prove Theorems 1 and 2 in terms of $w^*$ which is a function of the primitives.

For the last part of this section we will restrict ourselves to the special case where the probability of observing a worker's effort is proportional to the amount of time she is supervised i.e. $\pi(s) = s$. Assume that:

$$f(p) = p^\alpha$$
$$u(w, e) = w - e^\beta$$

Here $\alpha$ and $\beta$ are positive constants. Notice that the optimal contract requires $w(e, s) = \frac{e^\beta}{s}$. Then our necessary condition is satisfied if there exists some $e \in [0, 1]$ and $s \in [0, e]$ such that:

$$(e - s)^\alpha - \frac{e^\beta}{s} \geq 0$$

When $\alpha = 1$ this inequality is only satisfied if $\beta > 2$. More generally, when $\alpha$ and $\beta$ are integers, this inequality is only satisfied if $\beta > 2\alpha$.[22] This suggests that incentive problems limit firm size

[22] For example the inequality is satisfied when ($\alpha = 2$, $\beta = 4$) but not when ($\alpha = 3$, $\beta = 4$).

unless worker productivity is "sufficiently" greater than the disutility of effort: incentive problems force the owner to either spend time on supervision or pay a wage which is strictly greater than the disutility of effort. In both cases, the total cost of hiring a worker is greater than what it would be without incentive problems. So a worker needs to be productive enough to compensate for this additional cost.

# 5 Conclusion

It is generally accepted that firms have an incentive problem because employers and employees have different objectives. This often results in a loss of control - the employee's action and the employer's preferred action are not the same. Indeed, one can find many examples of organizational failures caused by incentive problems. Not surprisingly, there have been many papers studying the role of incentives within a firm. One open question is the extent to which incentive problems prevent a firm from growing larger.

Most of the existing literature finds that incentive problems limit firm size only when we make some very specific assumptions. This paper shows that that is not necessary. We find that firm size is limited even in the CW78 model if we allow for a non-linear production function. We then develop a very general model of supervision and wage incentives to establish necessary and sufficient conditions under which the optimal size of a firm is indeterminate.

Our model assumes that supervision is costly. So workers need to generate enough profit to compensate for the cost of supervising them. This is only possible when worker productivity is sufficiently greater than the disutility of effort. Our necessary condition concretizes this intuition. We show that incentive problems limit firm size unless a worker is able to generate positive net profit while also supplying at least as much supervision as he receives.

We prove our results using a novel optimization technique which is a significant contribution on its own. This technique allows us to obtain bounds on the profit generated by an infinitely large firm. It can be applied to any dynamic programming problem where the lack of a discount factor makes it impossible to employ the standard Bellman equation approach.

Our results show that incentive problems can limit firm size even in the absence of other inefficiencies like communication costs or incomplete contracts. We provide a foundation for re-evaluating the role of incentive problems as a constraint on firm growth. Because we make minimal assumptions, our results are widely applicable and have important implications for large organization of all kinds.

# References


Patrick Bolton and Mathias Dewatripont. The firm as a communication network. *The Quarterly Journal of Economics*, 109(4):809–839, 1994.

Guillermo A Calvo and Stanislaw Wellisz. Supervision, loss of control, and the optimum size of the firm. *Journal of political Economy*, 86(5):943–952, 1978.

Ronald H Coase. The nature of the firm. *Economica*, 4(16):386–405, 1937.

Saikat Datta. On control losses in hierarchies: a synthesis. *Rationality and Society*, 8(4):387–412, 1996.

Luis Garicano. Hierarchies and the organization of knowledge in production. *Journal of political economy*, 108(5):874–904, 2000.

Sanford J Grossman and Oliver D Hart. Implicit contracts under asymmetric information. *The Quarterly Journal of Economics*, pages 123–156, 1983.

Sanford J Grossman and Oliver D Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of political economy*, 94(4):691–719, 1986.

Oliver Hart and John Moore. Property rights and the nature of the firm. *Journal of political economy*, 98(6):1119–1158, 1990.

Bengt Holmström. Moral hazard and observability. *The Bell journal of economics*, pages 74–91, 1979.

Michael Keren and David Levhari. The internal organization of the firm and the shape of average costs. *The Bell Journal of Economics*, pages 474–486, 1983.

James A Mirrlees. The optimal structure of incentives and authority within an organization. *The Bell Journal of Economics*, pages 105–131, 1976.

Yingyi Qian. Incentives and loss of control in an optimal hierarchy. *The review of economic studies*, 61(3):527–544, 1994.

Roy Radner. The organization of decentralized information processing. *Econometrica: Journal of the Econometric Society*, pages 1109–1146, 1993.

Steven Shavell. Risk sharing and incentives in the principal and agent relationship. *The Bell Journal of Economics*, pages 55–73, 1979.

Masatoshi Tsumagari. *Incentives, monitoring and firm structure*. Boston University, 1999.

Oliver E Williamson. Hierarchical control and optimum firm size. *Journal of political economy*, 75 (2):123–138, 1967.

# Appendix

We prove the general versions of our results which are stated in Section 4. The versions stated in Section 2 are special cases.

*Proof of Existence.* Recall that $\psi(e, s)$ is the distribution of $w$ given some $e$ and $s$.

Define $m_{e,s}$ to be the measure on the probability space induced by $\psi(e, s)$. So $m_{e,s}(w)$ is the conditional probability of $w$ for a given value of $e$ and $s$.

The remainder of this proof assumes a fixed value for $s$. Where there is no ambiguity, we will omit the $s$ variable from our notation.

The additional assumption which we need is that for any integrable function $f$ and any sequence $e_n \to e$ the measure satisfies:

$$\int_w f m_e(dw) \geq \limsup_n \int_w f m_{e_n}(dw)$$

Notice that we assume the above inequality holds for a fixed function $f$. So our desired result is not immediate. Intuitively, the above assumption imposes a notion of upper semi-continuity on the measure.

We have already assumed that the domain of $e$ is compact (recall that effort is restricted to the unit interval). So a solution to the maximization problem is guaranteed if we can show that $\mathbb{E}_{w\sim\psi}[u(e, w)]$ is upper semi-continuous. That means for any $e_n \to e$ we have:

$$\int_w u(e, w) m_e(dw) \geq \limsup_n \int_w u(e_n, w) m_{e_n}(dw)$$

The right hand side of the inequality can be written as:

$$\begin{aligned}\limsup_n \int_w u(e_n, w) m_{e_n}(dw) = &\limsup_n \int_w u(e_n, w) \mathbb{1}\big[u(e_n, w) \leq u(e, w)\big] m_{e_n}(dw) \\ &+ \limsup_n \int_w u(e_n, w) \mathbb{1}\big[u(e_n, w) > u(e, w)\big] m_{e_n}(dw)\end{aligned}$$

Where $\mathbb{1}$ represents the indicator function. Notice that the first term satisfies:

$$\limsup_n \int_w u(e_n,w)\mathbb{1}\big[u(e_n,w)\le u(e,w)\big]m_{e_n}(dw) \le \limsup_n \int_w u(e,w)m_{e_n}(dw)$$
$$\le \int_w u(e,w)m_e(dw)$$

Where the second inequality comes from our additional assumption. Applying Fatou's Lemma to the second term gives:

$$\limsup_n \int_w u(e_n,w)\mathbb{1}\big[u(e_n,w)> u(e,w)\big]m_{e_n}(dw)$$
$$\le \int_w \limsup_n \Big[u(e_n,w)\mathbb{1}\big[u(e_n,w)> u(e,w)\big]m_{e_n}\Big](dw)$$
$$= \int_w 0\times \limsup_n \Big[m_{e_n}\Big](dw)$$
$$= 0$$

Combining everything gives:

$$\limsup_n \int_w u(e_n,w)m_{e_n}(dw) = \limsup_n \int_w u(e_n,w)\mathbb{1}\big[u(e_n,w)\le u(e,w)\big]m_{e_n}(dw)$$
$$+\limsup_n \int_w u(e_n,w)\mathbb{1}\big[u(e_n,w)> u(e,w)\big]m_{e_n}(dw)$$
$$\le \int_w u(e,w)m_e(dw)+0$$

This proves that $\mathbb{E}_{w\sim\psi}[u(e,w)]$ is upper semi-continuous and completes the proof of existence.

□

We start the proof of Lemma 1 by providing a formal statement of the claim.

**Lemma 1.** *Let $n$ be the number of workers supervised by a manager. Define $\pi_n$ to be the maximum profit possibly generated by the $n$ workers. There exists some $N$ such that $\pi_n < \pi_{n-1}$ whenever $n > N$.*

*Proof of Lemma 1.* Recall that workers have an outside option which gives them utility $\underline{u} > 0$.

The participation constraint requires that a worker's expected wage is at least $\underline{w} > 0$.

Define $\underline{e} = \min f^{-1}(\underline{w})$. Then $f(p)$ increasing means that a worker will generate negative profit whenever $e < \underline{e}$.

The proof is clearly trivial if any of the $n$ workers is exerting less than $\underline{e}$ effort. So we will assume that $e \ge \underline{e}$ for all workers.

Let $e$ be the total effort exerted by the manager. Let $e_i$ be the total effort exerted by worker $i$ and let $s_i$ be the supervision received by him.

Without loss of generality assume that $s_1 = \min_i s_i$. Then $s_1 \leq \frac{e}{n}$.

Recall that $w(e, s)$ is increasing in $e$:

$$\begin{aligned} f(e_1) - w(e_1, s_1) &\leq f(e_1) - w(\underline{e}, s_1) \\ &\leq f(1) - w(\underline{e}, s_1) \end{aligned}$$

But we know that $w(e, s) \to \infty$ as $s \to 0$. So:

$$\begin{aligned} n \to \infty &\implies \frac{e}{n} \to 0 \\ &\implies s_1 \to 0 \\ &\implies w(\underline{e}, s_1) \to \infty \end{aligned}$$

Then there exists some $N$ such that for every $n > N$:

$$w(\underline{e}, s_1) > f(1)$$

This shows that Worker 1 is generating negative profit when $n > N$. Clearly total profit can be increased by removing Worker 1. Thus $\pi_n < \pi_{n-1}$ for any $n > N$.

□

*Proof of Lemma 2.* Consider any firm with $N < \infty$ layers.

Assume layer $n$ has $k < \infty$ workers.

Lemma 1 tells us that there exists some $M \in \mathbb{N}$ such that no manager can supervise more than $M$ workers.

So it must be the case that layer $n + 1$ is finite with no more than $k \times M$ workers.

Recall that the firm has a single owner so layer 0 is finite. Thus, by induction, we know that each layer of the firm must be finite.

The finite union of finite sets is always finite. So any firm with a finite number of layers must have a finite number of workers.

□

*Proof of Lemma 3.* Select any worker in layer $N$ and call him worker $n$.

By definition, worker $n$ will have a manager who is in layer $N-1$. Call him worker $n-1$.

Continuing in this manner, construct a finite sequence of workers such that worker $i$ is in layer $i$ and is supervised by worker $i-1$. Terminate the sequence when layer 1 is reached.

Notice that this is precisely the definition of a branch. Furthermore, the firm is assumed to have $N$ layers so the constructed branch will have exactly $N$ workers.

□

*Proof of Theorem 1.* Define $M = f(e^* - s^*) - w(e^*, s^*) > 0$ and define $M_0 = \max_s f(1-s) + f(e^*) - w(e^*, s)$.

Notice that $M_0$ is the net revenue of a firm in which the owner employs exactly one worker and picks the optimal level of supervision to induce $e^*$ effort.

Consider a firm with two workers. Suppose both workers exert $e^*$ effort; Worker 2 is supervised by Worker 1 and receives $s^*$ supervision; Worker 1 is supervised by the owner. Then the firm has two layers (excluding the owner) and its total profit is given by:

$$f(1-s) + f(e^* - s^*) + f(e^*) - w(e^*, s) - w(e^*, s^*) = M_0 + M$$

The owner can continue adding single employee layers in this manner. The total profit of the $n$ layer firm will be given by $M_0 + (n-1)M$.[23]

Since $M > 0$, total profit is strictly increasing in $n$. Thus, the firm can grow infinitely large and generate infinite profit.

□

*Proof of Theorem 2.* Fix an infinitely large firm generating infinite profit.

Lemma 2 and Lemma 3 tell us that this firm must have at least one infinitely long branch. Pick any such branch.

Define $e_n$ to be the effort exerted by the worker in layer $n$ of the infinite branch and define $s_n$ to be the supervision received by him. Then $p_n = e_n - s_{n+1}$ is the level of productive effort exerted by the worker. Note that both $\{e_n\}$ and $\{s_n\}$ form infinite sequences.

[23] Layer 0 is the owner so $n$ is defined as the number of layers excluding the owner.

The total profit generated by this branch is given by:

$$\pi = \sum_{n=1}^{\infty} f(e_n - s_{n+1}) - w(e_n, s_n)$$

Define $(\overline{e}, \overline{s}) = \arg\max_{e,s} f(e-s) - w(e,s)$.[24]

Now consider the following maximization problem:

$$\begin{aligned} \underset{e,\, s,\, s'}{\text{maximize}} \quad & f(e - s') - w(e, s) \\ \text{subject to} \quad & s' \geq s \end{aligned}$$

Recall that $f$ is assumed to be increasing so this problem will be solved by $e = \overline{e}$ and $s = s' = \overline{s}$. But then the Kuhn-Tucker theorem says that there exists $0 \leq \lambda < \infty$ such that:

$$f(e - s') - w(e, s) + \lambda(s' - s) \leq f(\overline{e} - \overline{s}) - w(\overline{e}, \overline{s}) \tag{1}$$

Now consider any sequence $(e_t, s_t)$. Inequality (1) tells us that for every $t$:

$$\begin{aligned} f(e_t - s_{t+1}) - w(e_t, s_t) + \lambda(s_{t+1} - s_t) &\leq f(\overline{e} - \overline{s}) - w(\overline{e}, \overline{s}) \\ f(e_t - s_{t+1}) - w(e_t, s_t) - f(\overline{e} - \overline{s}) + w(\overline{e}, \overline{s}) &\leq \lambda(s_t - s_{t+1}) \end{aligned}$$

Summing up over all $t$ gives:

$$\begin{aligned} \lim_{T\to\infty} \sum_{1}^{T} [f(e_t - s_{t+1}) - w(e_t, s_t) - f(\overline{e} - \overline{s}) + w(\overline{e}, \overline{s})] &\leq \lim_{T\to\infty} \sum_{1}^{T} \lambda(s_t - s_{t+1}) \\ &= \lambda(s_1 - s_T) \\ &< \infty \end{aligned}$$

This shows that the revenue generated by any sequence $(e_t, s_t)$ is at most finitely greater than the revenue generated by the constant sequence $(\overline{e}, \overline{s})$.

We assumed that the firm as a whole is generating infinite profit. So it cannot be the case that the branch in question is generating infinite loss. But that requires $f(\overline{e} - \overline{s}) - w(\overline{e}, \overline{s}) \geq 0$. We know that $\overline{e} \in [0, 1]$ and $\overline{s} \in [0, \overline{e}]$ so setting $(\overline{e}, \overline{s}) = (e^*, s^*)$ completes the proof.

□

[24] If a maximum does not exist, we can replace $f(\overline{e} - \overline{s}) - w(\overline{e}, \overline{s})$ with $\sup_{e,s} f(e-s) - w(e,s)$.