Incentives, Supervision, and Limits to Firm Size

Ayush Gupta
Boston University

3 July 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 dis-utility of effort.

1 Introduction

Standard economic models treat firms as singular agents with well defined objectives. But firms, like all organizations, are made up of multiple individuals with competing interests. The literature on Principal-Agent models tells us that this typically creates incentive problems which result in the employer having limited control over her employees' actions (called "loss of control" in the literature).[1] Organization economists have long tried to establish whether this loss of control compounds as the size of a firm increases. If it compounds, incentive problems alone may impose a limit on the size of a firm.

The literature in this field dates back to Coase (1937). Coase proposed that firms exist to avoid transaction costs but cannot grow arbitrarily large because of decreasing returns to the entrepreneur function. Williamson (1967) formalized the notion that loss of control across successive hierarchical levels results in internal diseconomies of scale and ultimately limits the size of a firm. However, Williamson did not attempt to model the source of this loss of control, assuming instead that "only a [exogenous] fraction $0 < \hat{\alpha} < 1$ of the intentions of a superior are effectively satisfied by a subordinate."

This paper builds on Calvo and Wellisz (1978) (hereafter "CW78") which endogenized $\hat{\alpha}$ using a model of supervision and wage incentives. They proved that, in their model, 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. Our paper starts by showing that their result rests on strong assumptions which may not be justified in all situations. For example, CW78 assumed that the profit generated by an individual worker is a linear function of his effort. If, instead, one assumes a concave function (which may be more appropriate when worker fatigue is a concern), we show that incentive problems can limit firm size.

We build a general model of incentive problems within a firm. Effort is costly so workers shirk on the job unless supervised by their manager. This forces managers to split their time between productive work (which generates profit) and supervisory work (which does not generate profit). The main contribution of this paper is to establish necessary and sufficient conditions for firm size to not be limited. These conditions state that incentive problems do not constrain the growth of a firm if it is profitable to reorganize the firm as a stationary hierarchy - a hierarchy in which each worker exerts the same level of productive and supervisory effort. Note that we do not claim that stationary hierarchies are optimal, nor do we restrict our analysis to such firms. Moreover, we make minimal assumptions about the production technology, the nature of supervision, or the type of employment contracts. As a result, our model generalizes much of the existing literature and our

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

results are widely applicable.

Our necessary and sufficient conditions are only satisfied when the productivity of effort is "sufficiently" greater than the dis-utility of effort.[2] That is because incentive problems necessitate supervision. Supervision is costly so workers need to generate enough profit to compensate for the cost of supervising them. That means that a worker's production needs to be sufficiently greater than his wage, which is proportional to the dis-utility of effort. In particular, firms may be limited to a finite number of workers when the productivity of effort is greater than its dis-utility, but the difference is not large enough to satisfy the necessary condition. Intuitively, the requisite difference is proportional to the severity of the incentive problem. It decreases when the monitoring technology becomes more effective or the firm is able to offer a more efficient contract.[3] An additional corollary is that firms can grow larger when productivity increases.

The significance of our result is that it shows that incentive problems *alone* may limit the growth of a firm, under a reasonable relaxation of the CW78 assumptions. To illustrate this point, we present a simple example of a firm which exhibits constant returns to scale.[4] We know that, under standard assumptions (with no incentive problem), the optimal size of this firm is indeterminate. We introduce incentive problems à la CW78 and show that it is still profitable for our hypothetical firm to hire a positive number of workers. However, it is no longer profitable for the firm to grow infinitely large. Incentive problems limit the firm to a positive but finite number of workers.

The proof of our necessary condition constitutes a significant contribution it 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.

This paper contributes to a large literature on the role of incentives in organizations. Empirical papers include Prendergast (1999); Shearer (2004); Cadsby et al. (2007); Aghion et al. (2014); Friebel et al. (2017); Dessein et al. (2022); Antón et al. (2023); Christensen et al. (2023); Englmaier et al. (2024) and many others.

On the theoretical side, Keren and Levhari (1983), Fumas (1993), Qian (1994), McAfee and McMillan (1995); Datta (1996), Tsumagari (1999); Meagher (2003); van den Brink and Ruys (2008);

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

[3]A contract is more efficient if it induces a higher level of effort for the same wage.

[4]Meaning that output increases linearly in the number of workers.

Hart and Holmstrom (2010); Galashin and Popov (2016); Chen (2017); and Hiller (2021) all study incentive problems in the context of firm size.

The existing literature is largely consistent with the CW78 result. It finds that incentive problems do not limit firm size unless one assumes a specific production function, introduces coordination costs, or makes other assumptions which limit the applicability of the results. For example, Qian (1994) assumes 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.

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 model, 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 firm with one owner who wants to maximize profit. The owner can operate alone or hire from an infinite pool of identical workers. The profit generated by each worker is given by $f(p)$ where p is the worker's level of productive effort. We assume that f is increasing, bounded, and that $f(0) = 0$. We do not make any additional assumptions about the slope or shape of f.[5]

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.[6] These assumptions ensure that, in the absence of incentive problems, the firm exhibits constant returns to scale i.e. total production increases linearly in the number of workers. We focus on constant returns to scale instead of decreasing returns to scale because the latter limits the optimal size of a firm even in the absence of incentive problems. Such an assumption would defeat our objective which is to show that incentive problems can limit firm size in situations where the optimal size is otherwise infinite.[7]

[5] The CW78 model assumes that f is linear.

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

[7] It is possible that incentive problems can limit firm size even when we have increasing returns to scale. However, studying such firms requires a substantially different framework and is beyond the scope of this paper.

The von Neumann-Morgenstern utility index of each worker is given by $U = u(w, e)$ 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 that u is increasing in wage and decreasing in effort with $u(0, e) \leq 0$ for all e. We also assume that all workers have a positive outside option which gives them $\underline{u} > 0$ utility.[8]

It is obvious that a worker who is offered a fixed wage[9] 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 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 5 relaxes this assumption and allows for any arbitrary form of supervision.

We assume effort-contingent contracts (rather than output-contingent contracts) in this section to stay consistent with CW78 and a majority of the existing literature. This assumption is typically justified on the grounds that it is often easier for the owner to verify effort rather than output. We relax this assumption in Section 5.

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 $\bar{w}$. The worker receives $\bar{w}$ if his effort level is not observed. If the worker's effort is observed, he receives $e\bar{w}$ where e is his level of effort.[10] Given some $\bar{w}$ and s, the worker picks e to maximize his expected utility:

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

Or he picks his outside option if $U < \underline{u}$. For simplicity, assume that there is a unique $e(\bar{w}, s)$ which maximizes the worker's utility (whenever his participation constraint is satisfied).[11] Uniqueness

[8] If $\underline{u} = 0$ it is possible for a 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.

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

[10] 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 any arbitrary contract.

[11] This will be true whenever U is continuous and strictly quasi-concave in e. This is not necessary for our results but we will maintain this assumption throughout the paper.

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

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

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 dis-utility of effort.
- $w(e, s) \to \infty$ as $s \to 0$ because unsupervised workers receive a fixed wage and never exert any effort.
- $w(e, s) \geq \underline{\mathrm{w}} > 0$ for all e because workers have an outside option which gives them $\underline{\mathrm{u}} > 0$ 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 effort and supervisory effort . 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.

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 effort 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.

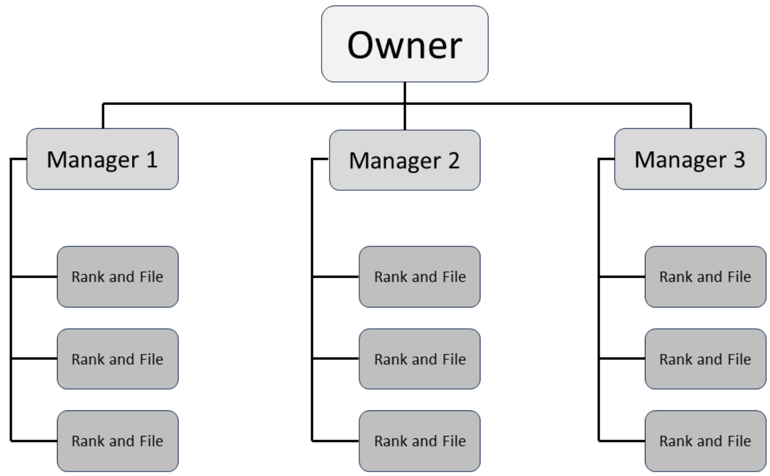

Figure 1: Example Firm Hierarchy

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. 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:

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

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

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 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

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

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

Theorem 1 proves a sufficient condition for incentive constraints to not limit firm size. Theorem 2 proves that a slightly weaker condition is necessary if we make the additional assumption that $f(p)$ and $w(e, s)$ are continuous over their entire domain.

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. As a result, firm size is limited unless workers can compensate the firm for this additional cost. The necessary condition serves to specify the requisite level of compensation. One interpretation is that workers need to supply at least as much supervision

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

as they receive. Then we can say that the net supervision received by workers is zero so it is enough for them to generate positive profit.

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 does not supervise anyone else. The left hand side represents the decrease in the worker's production if he exerts s^* supervisory effort. It can be interpreted as the "value" of the supervision received by the worker.[14] Then 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 stationary hierarchy and remain profitable. In practice, a firm might find it optimal 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

[14] Under this definition the value of supervision received by a worker can be different from the opportunity cost of supervising said worker.

[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.

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. 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:

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

Suppose that employment contracts follow the CW78 model i.e. workers are paid $\bar{w}$ if their effort is not observed and they are paid $e\bar{w}$ if their effort is observed. Then we know that workers will exert $e = \frac{1}{2}s\bar{w}$ and $w(e, s) = \frac{2e}{s} - 2e(1 - e)$.

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 know 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 $\frac{2e}{s} - 2e(1 - e) \in [\frac{3}{2}, \infty)$. On the other hand, we 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. To see how this might happen, suppose that the owner hires a worker who receives s supervision and exerts e effort. Call him Worker 1. For Worker 1 to profitably supervise another worker, there must be some $e' \in [0, 1]$ and some $s' \in [0, e]$ such that:

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

However, no such e' and s' exists if $e < \frac{1}{3}$. That prevents a second worker from being hired because Worker 1 exerting $e > \frac{1}{3}$ is never profitable.

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 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. An alternate interpretation is that the incentive problem is too severe. This interpretation comes from the fact that the necessary condition may be satisfied if Worker 1 can be induced to exert more effort for the same wage (that is equivalent to the incentive problem becoming less severe).

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:

[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.

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

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 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 misaligned incentives, 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.

In the absence of incentive problems, constant returns to scale are a natural consequence of the fact that we assume individual worker output is aggregated linearly and that there is no complimentarity in worker effort. 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 impose a limit on 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 relaxes many of the assumptions made in Section 2. We start with the employment contracts. Previously, we were assuming that the employer is restricted to a specific contract such that the wage received by a worker who exerts effort e and receives supervision s is given by:

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

We now allow the owner to offer any arbitrary contract as long as it satisfies the limited liability constraint. So $\omega(e, s)$ is allowed to be an arbitrary function.[17]

[17] We define $\omega(e, s)$ as the expected wage of a worker who exerts effort e and receives supervision s. This is different from $w(e, s)$ which is the expected wage that the owner needs to pay to induce effort e when the level of supervision is s.

The expected utility maximizing workers will solve the following problem:

$$U = \max_{e} \mathbb{E}_{\omega}[u(e, \omega(e, s))]$$

Allowing ω to be an arbitrary function also helps us relax our previous assumption about the nature of supervision. We no longer assume that the level of supervision received by a worker is equal to the probability of her effort being observed. Instead, we allow for any arbitrary relationship between the level of supervision and what is observed about a worker's effort level. For example, it might be the case that a manager can perfectly observe effort if the level of supervision is greater than some $\underline{s} < 1$. This possibility (and many others) are allowed under our approach and are captured by the function ω.

Finally, notice that the function ω is silent about what is being observed or how the wage is being determined. In particular, consider an output contingent contract. There is no uncertainty in our model so output can be expressed as a function of effort. Then any output contingent contract is equivalent to an effort contingent contract. Of course, if a worker spends part of his effort on supervising others, it is not obvious how an output contingent contract should reward effort. For example, the employer may choose to compensate a manager based on the output of his subordinates. However, as long as the subordinates' output is a function of the manager's effort, such a contract can be expressed as an effort contingent contract.

Recall that workers exert the level of effort which maximizes their own utility. So, given some level of supervision, worker effort depends only on the relationship between effort and wage. In particular, if the wage is not a function of effort, workers will exert zero effort. That is why, under any relevant employment contract, the wage can be expressed as a function of effort.[18]

To summarize, there are two key assumptions underpinning our model. The first is that a worker's wage depends on his level of effort. This may be directly (in the case of effort contingent contracts) or indirectly (in the case of output contingent contracts). The second key assumption is that enforcing employment contracts requires supervision, which is costly. We do not need to make assumptions about how the wage depends on effort or about how costly supervision is.

As before, we focus on the expected wage function $w(e, s)$ which will depend on the nature of supervision and the type of employment contract used by the owner. We continue to assume the following:

- Worker utility is decreasing in effort. So $w(e, s)$ is increasing in e because workers need to be compensated for the dis-utility of effort.

[18] A relevant employment contract is any contract which induces a positive level of effort.

- Some positive level of supervision is necessary to incentivize workers. So $w(e,s) \to \infty$ as $s \to 0$.
- Workers have a positive outside option which gives them $\underline{\text{u}} > 0$ utility. So $w(e,s) \geq \underline{\text{w}} > 0$ for all e.[19]

We further assume that the individual production function and the expected wage function are continuous over their entire domain.[20] This allows us to prove all of our results as presented in Section 3. 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:*

$$f(e^* - s^*) > w(e^*, s^*)$$

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:*

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

We note that Theorems 1 and 2 establish conditions on the expected wage function which is endogenously determined.[21] 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 supervision technology. To do otherwise would result in a different condition for every possible combination of utility function, employment contract, and supervision technology. This is both infeasible and uninformative.

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

[20] The expected wage function is endogenously determined and need not be continuous in general. However, it can be mapped to the primitives and this assumption can be stated as an assumption on the utility function. We explain this later in the section.

[21] In particular $w(e,s)$ depends on the utility function, the supervision technology, and the choice of employment contract.

Moreover, the utility function and the supervision technology are both exogenous, so there is 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 owner's perspective. Under this contract the owner specifies a minimum acceptable level of effort $\underline{\mathrm{e}}$. The worker receives wage $\bar{w}$ if he exerts $e \geq \underline{\mathrm{e}}$. If the worker is observed to exert $e < \underline{\mathrm{e}}$ effort, he receives zero wage. Then the worker exerts effort $\underline{\mathrm{e}}$ if and only if:

$$u(\bar{w}, \underline{\mathrm{e}}) > (1 - \pi(s))u(\bar{w}, 0)$$

and

$$u(\bar{w}, \underline{\mathrm{e}}) > \underline{\mathrm{u}}$$

Here $\pi(s)$ is the probability that the worker's effort is observed. This is a function of the level of supervision received by the worker and the supervisory technology used by the firm. Then, given some level of supervision s and desired level of effort e, the optimal wage w^* is given by:

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

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

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

In equilibrium, the worker will always exert effort e and receive wage w^*. The expected wage function is thus given by $w(e, s) = w^*$.[22] We can now state Theorems 1 and 2 using only 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. $\pi(s) = s$ (this is the CW78 assumption). Assume that:

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

Here α and β 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 α and β are integers,

[22]This means that $w(e, s)$ will be continuous whenever u and π are continuous.

this inequality is only satisfied if $\beta > 2\alpha$.[23] This suggests that incentive problems limit firm size unless worker productivity is "sufficiently" greater than the dis-utility of effort: incentive problems force the owner to either spend time on supervision or pay a wage which is strictly greater than the dis-utility of effort. In both cases, the total cost of hiring a worker is greater than what it would be *sans* 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 dis-utility 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.

Because we make minimal assumptions, our results are widely applicable. They allow the literature to re-examine incentive problems as one of the drawbacks of a vertically integrated firm, preventing indiscriminate consolidation in the economy. Our results also have important implications for the

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

study of bureaucracies and other large organizations in the political sphere.

References

Phillipe Aghion, Nicholas Bloom, and John Van Reenen. Incomplete contracts and the internal organization of firms. *The Journal of Law, Economics, & Organization*, 30(suppl_1):i37–i63, 2014.

Miguel Antón, Florian Ederer, Mireia Giné, and Martin Schmalz. Common ownership, competition, and top management incentives. *Journal of Political Economy*, 131(5):1294–1355, 2023.

C Bram Cadsby, Fei Song, and Francis Tapon. Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of management journal*, 50(2):387–405, 2007.

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.

Cheng Chen. Management quality and firm hierarchy in industry equilibrium. *American Economic Journal: Microeconomics*, 9(4):203–244, 2017.

Peter Christensen, Paul Francisco, and Erica Myers. Incentive-based pay and building decarbonization: Experimental evidence from the weatherization assistance program. Working Paper 31322, National Bureau of Economic Research, June 2023.

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.

Wouter Dessein, Desmond Lo, and Chieko Minami. Coordination and organization design: Theory and micro-evidence. *American Economic Journal: Microeconomics*, 14(4):804–843, 2022.

Florian Englmaier, Stefan Grimm, Dominik Grothe, David Schindler, and Simeon Schudy. The effect of incentives in nonroutine analytical team tasks. *Journal of Political Economy*, 132(8): 000–000, 2024.

Guido Friebel, Matthias Heinz, Miriam Krueger, and Nikolay Zubanov. Team incentives and performance: Evidence from a retail chain. *American Economic Review*, 107(8):2168–2203, 2017.

Vincente Salas Fumas. Incentives and supervision in hierarchies. *Journal of Economic Behavior & Organization*, 21(3):315–331, 1993.

Mikhail Galashin and Sergey V Popov. Teamwork efficiency and company size. *The BE Journal of Theoretical Economics*, 16(1):337–366, 2016.

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

Oliver Hart and Bengt Holmstrom. A theory of firm scope. *The quarterly journal of economics*, 125(2):483–513, 2010.

Tobias Hiller. Hierarchy and the size of a firm. *International Review of Economics*, 68(3):389–404, 2021.

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.

R Preston McAfee and John McMillan. Organizational diseconomies of scale. *Journal of Economics & Management Strategy*, 4(3):399–426, 1995.

Kieron J Meagher. Generalizing incentives and loss of control in an optimal hierarchy: the role of information technology. *Economics Letters*, 78(2):273–280, 2003.

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

Canice Prendergast. The provision of incentives in firms. *Journal of economic literature*, 37(1): 7–63, 1999.

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

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

Bruce Shearer. Piece rates, fixed wages and incentives: Evidence from a field experiment. *The Review of Economic Studies*, 71(2):513–534, 2004.

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

René van den Brink and Pieter HM Ruys. Technology driven organizational structure of the firm. *Annals of Finance*, 4:481–503, 2008.

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 5. The versions stated in Section 3 are special cases.

Proof of Lemma 1. First define $e_{min} = \inf e \in [0,1]$ such that $f(e_{min}) \geq \underline{\text{w}}$. Next consider a manager who exerts effort e and supervises N workers. Without loss of generality, define $s_1 = \min\{s_1 ... s_N\}$ where s_i is the level of supervision received by worker i. We know that $s_1 \leq \frac{e}{N}$ so $s_1 \to 0$ as $N \to \infty$.

Then there exists some M such that $w(e_{min}, s_1) > f(1)$ for all $N > M$. So $f(e_1) - w(e_1, s_1) \leq f(1) - w(e_{min}, s_1) < 0$ for every $e_1 \geq e_{min}$.

Moreover, $f(e_1) - w(e_1, s_1) \leq f(e_1) - \underline{\text{w}} < 0$ for every $e_1 < e_{min}$. □

Proof of Lemma 2. Consider a firm with $N < \infty$ layers. Suppose 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 kM 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.

Recall that 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 we can construct a sequence of workers such that worker i is in layer i and supervises worker $i+1$. 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 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.

Now suppose that the owner adds another layer with exactly one worker. We know that $M > 0$

means that $e^* > s^*$ which means that it is feasible for the worker in layer 1[24] to supervise the worker layer 2 for s^* units of time. Suppose that the owner picks this (possibly sub-optimal) arrangement and offers a wage which induces e^* effort from the worker in layer 2. Then the total profit of this firm 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 $M_0 + (n-1)M$.[25] Since $M > 0$ we know that total profit is strictly increasing in n so the firm can reach infinite size and profit. □

Proof of Theorem 2. Assume that we have 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. Consider 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 each worker. Note that both $\{e_n\}$ and $\{s_n\}$ form infinite sequences.

The total profit generated by this branch is given by:

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

Now define $(\bar{e}, \bar{s}) := argmax_{e,s} f(e-s) - w(e,s)$.

Note that f and w are upper semi-continuous functions defined on a compact set so the maximum always exists. Recall that f is assumed to be increasing so $(\bar{e}, \bar{s})$ will maximize $f(e-s') - w(e,s)$ subject to $s' \geq s$. But then the Kuhn-Tucker theorem says that there exists $\lambda \geq 0$ such that:

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

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

$$f(e_t - s_{t+1}) - w(e_t, s_t) + \lambda(s_{t+1} - s_t) \leq f(\bar{e}-\bar{s}) - w(\bar{e}, \bar{s})$$
$$f(e_t - s_{t+1}) - w(e_t, s_t) - f(\bar{e}-\bar{s}) + w(\bar{e}, \bar{s}) \leq \lambda(s_t - s_{t+1})$$

Summing up gives:

[24] Recall that the owner is always assumed to be Layer 0.
[25] Layer 0 is the owner so n is defined as the number of layers excluding the owner.

$$\sum_{1}^{T} [f(e_t - s_{t+1}) - w(e_t, s_t) - f(\bar{e} - \bar{s}) + w(\bar{e}, \bar{s})] \leq \sum_{1}^{T} \lambda(s_t - s_{t+1})$$
$$= \lambda(s_1 - s_T)$$
$$\leq M$$

Where M is some finite number. 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 $(\bar{e}, \bar{s})$.

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. That requires $f(\bar{e} - \bar{s}) - w(\bar{e}, \bar{s}) \geq 0$. We know that $\bar{e} \in [0, 1]$ and $\bar{s} \in [0, \bar{e}]$ so setting $(\bar{e}, \bar{s}) = (e^*, s^*)$ completes the proof. $\square$