Diseconomies of Scale in Employment Contracts

August 14, 1992

Eric Rasmusen and Todd Zenger

Published: Journal of Law, Economics and Organization (June 1990), 6: 65-92 *Abstract*

We find that small teams can write more efficient incentive contracts than large teams when agents choose individual effort levels but the principal observes only the joint output. This result is helpful in understanding organizational diseconomies of scale and is consistent with both existing evidence and our own analysis of data from the Current Population Survey. Our modelling approach, similar to classical hypothesis testing, is of interest because we need not derive the optimal contract to show the advantage of small teams.

Rasmusen: John M. Olin Visiting Assistant Professor, Center for the Study of the Economy and the State and Graduate School of Business, University of Chicago 60637, and UCLA AGSM. Phone: (312) 702-9478. Fax: (312) 702-0458. Bitnet: Fac_ras@gsbacd.uchicago.edu.

Zenger: School of Business and Management, Pepperdine University, 3rd floor, 400 Corporate Point, Culver City, Cal. 90230. Phone: (714) 941-0329. Fax: 213-568-5727.

of Business, BU 456, 1309 E 10th Street, Bloomington, Indiana, 47405-1701. Office: (812) 855-9219. Fax: 812-855-3354. Erasmuse@indiana.edu. Php.indiana.edu/∼erasmuse.

I. Introduction.

The Problem of Firm Size.

A central question in industrial organization is what determines the size of firms. This question is closely linked to a central assumption in microeconomics generally: that "managerial diseconomies of scale" limit the size of firms. If such diseconomies of scale did not exist, many industries would be natural monopolies, because average cost declines with output for any technology with a fixed cost and constant marginal cost. This was noted by Sraffa (1926) and Kaldor (1934), who concluded that perfect competition was unrealistic. Modern textbook theory, rejecting the premise of constant marginal cost, notes that if marginal cost rises with output, then a large firm is inefficient and a competitive market is feasible. The combination of rising marginal cost and a fixed cost generates the Vinerian U-shaped average cost curve, not natural monopoly.

The commonly given reason for increasing long-run marginal cost, a reason mentioned as far back as Kaldor (1934), is that a larger firm is harder to manage. Why this should be so is not entirely clear. Indeed, Alfred Marshall believed that there were managerial *economies* of scale, which would aggravate the problem noted by Kaldor and Sraffa.[1] Schumpeter also seems to have believed in managerial economies of scale, because "monopolization may increase the sphere of influence of the better, and decrease the sphere of influence of the inferior, brains."[2] Modern casual empiricism, if not that of Alfred Marshall, suggests that large organizations suffer from diseconomies. Not only is there no tendency towards monopoly in most unregulated markets, but in most industries one observes a variety of firm sizes.[3] Coase (1937) has argued that optimal firm size is determined by a comparative assessment of the costs of internalizing additional transactions and the costs of market transactions. At some point the cost of organizing another transaction within the firm becomes greater than the cost of transacting in the open market. But Coase merely provided a general framework, not a reason why transactions costs should rise as the firm becomes larger. As he notes in a 1988 article, why the marginal costs of internal organization should rise with increased size remains unexplained.

Recently, several theorists have developed arguments to explain organizational diseconomies of scale. Williamson (1985, Chapter 6) argues that common ownership of successive stages of production, unlike separate ownership, creates incentives for managers governed by incentive contracts to misutilize assets and opportunistically manipulate accounting data. Consequently, joint ownership of successive production stages requires costly supplemental monitoring to enforce incentive contracts. These enforcement costs and residual opportunistic behavior render incentive contracts inefficient, thereby discouraging joint ownership. Milgrom and Roberts (1988a,b) and Holmstrom (1988) argue that organizations, unlike markets, must contend with "influence costs"— costs that arise whenever individuals try to influence decisions to their private benefit and whenever organizations impose mechanisms to control this behavior. But the above arguments, although helpful in specifying the costs of internal organization, do not address why these or other costs of internal organization would necessarily increase with the size of the firm.

Diseconomies of scale also arise in organizations other than firms. Sugden (1986, chapter 5) and, more formally, Boyd and Richerson (1988) find that increasing the size of a group makes evolutionary development of a convention— a coordination rule such as driving on the right— more difficult. A convention works better when used more widely, and moving randomly from discoordination to coordination is easier in a smaller group. Farrell and Lander (1989) use a team model to look at the distinction between an individual's effort on his own behalf and on the team's behalf under certainty. They find for a particular contractual form that selfish effort increases and team effort decreases in team size. More directly related to firm size, Robert Frank (1985) in *Choosing the Right Pond* argues that if managers are willing to sacrifice monetary compensation in exchange for being top man in their firms, then small firms could outcompete large firms.

Bendor & Mookherjee (1987) analyze the effect of group size on cooperation in an infinitely repeated Prisoner's Dilemma. They note that as the group size becomes larger, the free rider problem increases. For a given discount rate, cooperation cannot be supported beyond a critical group size that depends on the discount rate. This has the same flavor as work on the size of cartels, such as Stigler's 1964 model of a cartel that tries to detect cheaters

using a noisy variable whose mean depends on whether cheating occurs. In his model, a firm only knows how many customers do not return to it, so each firm sees a different statistic. As the number of firms increases, detecting a price-cutter becomes more difficult. The problem has been taken up again in the "trigger strategy" literature, where detection triggers dissolution of the cartel. Porter (1983) comes to a result similar to Stigler's: a cartel with more members has a shorter expected lifetime.

Our Approach.

We hope to shed some light on the issue of organizational diseconomies of scale from a new viewpoint: thinking of the organization as a team of agents who may differ in effort or ability. We will show why teams incur diseconomies of scale and why teams can coexist with different management styles. Our argument will be that the small team can more efficiently construct incentive contracts to induce high effort and attract talented workers, an argument that applies to any organization generating a joint output when the inputs of individual members are difficult to assess.

We will use an agency model similar to Holmstrom (1982) in which total team output is observable but individual contributions are not. Holmstrom focuses on the *manager's* role in ensuring optimal effort. Even without a manager the team faces the difficulty of detecting low effort, but the manager's presence expands the space of contracts by enforcing drastic punishment if output is low. In a world of certainty this allows the first-best to be achieved. The Holmstrom model by itself thus does not explain diseconomies of scale; on the contrary, one may draw from it the surprising implication that a properly designed contract can avoid them.

We will assume that contracts are enforceable in that managers will always impose the punishments stipulated by the contract, and focus instead on team size under uncertainty, where low output might be due to random noise instead of shirking. The manager must still deter shirking by the threat of a teamwide punishment if output is below a chosen threshold. We will examine the case where the optimal contract is costly: even if inflicting the punishment incurs deadweight loss, some mistaken punishment occurs because of the noise, and the first-best cannot be achieved. We will show that

the costs associated with the optimal contract increase with team size. This result will be obtained in a way that avoids the difficult task of characterizing the optimal contract. Our argument will apply both to moral hazard (when effort is variable) and adverse selection (when ability is variable). The implication is that large teams must be content with low effort and ability or else use ability testing and effort monitoring instead of output-based contracts.

We believe that our results on teams are relevant to thinking about organizational diseconomies of scale. In the case of small firms, our results apply directly, since the firm may be nothing more than a small team of production workers. Large firms, however, are commonly comprised not of one large team, but of many subunits. Despite this, it is relatively uncommon for any significant element of compensation to be explicitly tied to subunit output. Why don't large firms simply replicate small-firm incentives by attaching compensation to subunit performance? Presumably, these subunits are not teams in our sense, with a clearly observable team output. Rather, the output of a subunit depends on the inputs of other subunits and management. Consequently, breaking into small subunits does not permit the large firm to replicate the contracting advantage of the small firm.[4]

Our model also has application to management teams. It may well be that the individual output of production workers is fully observable, but the board of directors cannot distinguish the individual contributions of the top five executives. In this case, our model predicts that firms with fewer top executives could construct contracts which attract greater talent at the executive level. If the number of executives limits the size of the firm, then only relatively small firms will employ high-ability executives and use incentive contracts.

The teams approach is important because it complements an older idea: the firm cannot simply replicate divisions, because it can have only one chief executive, who runs into diminishing returns when used intensively. This is the idea extended and modelled by the literature starting with Williamson (1967) and continuing in, for example, Keren and Levhari (1983) and McAfee and McMillan (1989a). Williamson constructs a model with an exogenous "span of control," an exogenous ratio of agents to managers at each level of the firm. Since managers are not directly productive, this assumption makes

average costs increase in output. The interesting questions in Williamson's paper concern the optimal number of levels in the firm hierarchy, however, rather than why large firms are inefficient, since the inefficiency is a direct assumption about the management technology. The span-of-control approach asks how to optimize the firm's structure given organizational diseconomies of scale, but it does not inquire into their origins. Our paper suggests that one reason for span-of-control problems is the increasing cost of extending incentive contracts as span increases.

The single-executive explanation for why replication is ineffective in removing managerial diseconomies of scale is satisfactory only if we take it for granted that a firm must have only one chief executive. But why not replace the chief executive with a management team, increasing the number of executives as the amount of necessary supervision increases? We will show that using a team is costly, and it is more costly the larger the team. Hence, making the management team larger by horizontal expansion of the hiearchy's top level is not a costless substitute for increasing the number of hierarchical levels by vertical expansion.

The teams approach may also explain why vertical expansion of the hierarchy with monitoring cannot be replaced by vertical incentive contracts. The individuals we model as a team need not be at the same hierarchical level. The reasoning applies even if one of the team members is a boss, whose ability is twice as high as an ordinary member and who is paid a double share, but whose effect on team output cannot be untangled from the effect of his subordinates. The results hold *a fortiori*, since low effort by the boss is as easy to detect as low effort by two ordinary members with perfectly correlated errors. All that is required is that the double importance of the boss is common knowledge. Thus, our model in its broadest interpretation suggests that as the number of individuals contributing to an observable output increases, the costs of offering incentive contracts rise regardless of the individuals' hiearchical location within the organization.

Sections II and III of this article present the model and show why contract costs increase with team size. Section IV discusses the applicability of the results to models of adverse selection, in which smaller firms can offer lower-cost contracts to attract high-ability agents while deterring the low-ability.

Section V lays out the empirical implications of the model and compares them with evidence from the 1979 Current Population Survey.

II. The Model.

A principal employs n agents to make up a team. The agents are identical, with a utility function $U(w, b, e)$ that is increasing in the wage w and decreasing in the punishment b and the effort e. Effort takes one of two levels, $\overline{q}_h$ or $\overline{q}_l$, where $\overline{q}_h > \overline{q}_l$. Agent i's contribution to output is the sum of his effort and a random disturbance:

$$q_h = \overline{q}_h + \varepsilon_i \quad \text{or} \quad q_l = \overline{q}_l + \varepsilon_i. \tag{1}$$

The disturbances ε_i are identically distributed with mean zero according to a multivariate normal distribution, and they may be either independent or positively correlated. If the disturbances are independent, the random influence is at the level of the individual agent; if they are perfectly correlated, it is at the level of the team. We use the normal distribution so that total team output follows the same distribution regardless of team size. This allows output to range from negative infinity to positive infinity, so some readers may wish to interpret the variable as some other measure of performance besides output (e.g. profit).

Competing teams offer contracts to the agents, who choose the contract that yields the greatest expected utility. After the agents choose contracts and efforts, nature chooses the values of the disturbances and the team outputs appear. The principals then pay or penalize the agents according to the terms of the contracts.

Individual effort and output are prohibitively costly for a principal to determine, so his contract must rely solely on the level of team output, denoted Q_{obs}. The principal is risk neutral and seeks to maximize profit, the expected value of $(Q_{obs} - nw)$. The simplest kind of contract ignores Q_{obs} and pays a fixed wage of $w = \overline{q}_l$, which elicits low effort and produce an expected profit of zero. If agents are sufficiently averse to risk and effort, this is the efficient outcome under asymmetric information— incentives for high effort are too costly. We will assume the opposite—that high effort is more efficient, even under asymmetric information— for the remainder of this article.

We will also assume that welfare under asymmetric information is not as high as it would be under full information— that is, the incentives necesssary

to induce high effort do generate costs. This is important because if neither large nor small teams incur costs, there is no cost difference between them. The assumption that incentives are costly can be justified by various combinations of primitive assumptions. If incentives require variability, and the agents are risk averse, then any variability in wages is costly. If the agents are risk neutral then wage variability does not matter, but if the principal is constrained in the wage contracts he can write he may be forced to use dissipative penalties. But the agent's utility function will matter only because of the two requirements that (a) high effort is efficient under asymmetric information, and (b) the asymmetry of information has no costless remedy.

The principal wishes to design a contract that supports a Nash equilibrium in which every agent exerts high effort. Since behavior under the Nash equilibrium concept presumes that each player calculates whether his unilateral deviation from equilibrium behavior is profitable, the willingness of one or more agents to switch would break the equilibrium. The principal's problem is therefore to make it unprofitable for even a single agent to choose low effort. If all n agents in a team choose high effort, output is the random variable

$$Q_{n,H} = \sum_{i=1}^{n} (\overline{q}_h + \varepsilon_i), \tag{2}$$

which is the equilibrium output. If all but one of the n agents choose high effort, output is the random variable

$$Q_{n,L} = \overline{q}_l + \varepsilon_1 + \sum_{i=2}^{n} (\overline{q}_h + \varepsilon_i), \tag{3}$$

which is the deviation output.

Testing for Shirkers.

Rather than directly attacking the problem of optimal contract design, we will start with the subproblem of detecting a low-effort agent. Under classical hypothesis testing, we establish a null hypothesis,

"H_0: All agents chose high effort, "

and an alternative hypothesis,

"H_1: At least one agent chose low effort."

H_1 is a little more general than is required here, since for Nash equilibrium all that is necessary is to test for one agent choosing low effort. A strategy combination is a Nash equilibrium if no player has any incentive to unilaterally deviate from his strategy. In this case, all agents will choose high effort in the proposed equilibrium, so we must test to see whether a single agent can benefit by deviating and choosing low effort. Whether several players might benefit by forming a coalition and simultaneously deviating is irrelevant to whether the strategy combination is a Nash equilibrium. But in this particular model, the Nash equilibrium concept is not restrictive. A test that detects deviation by a single low-effort agent would even more easily detect deviation by several agents who chose low effort simultaneously.

The principal is concerned with two types of errors:

Type I Error: Rejecting H_0, when it should be accepted.
(false rejection, associated with a low significance level)

Type II Error: Accepting H_0, when it should be rejected.
(false acceptance, associated with low power)

The principal wishes to avoid both types of error. Equivalently, he wants the detection test to have high values for (1) its *power* (the probability that a low-effort agent will be detected when present) and (2) its *significance level* (the probability of avoiding false detection).[5] If the power is high, the firm's probability of punishing a shirking agent is high and the agents fear to shirk. If the significance level is high, false detection is rare and agents need not be paid a large premium as compensation for expected accidental punishments. Both the power and the significance level are determined by the particular test.

Economists are used to trading off the levels of desirable characteristics. Here, however, the principal just desires a power that deters shirking. Higher levels of power are no better, so the lexicographic approach of classical hy-

pothesis testing is appropriate. The classical statistician chooses the significance level for a test and then maximizes the power given that significance level. Here, the principal chooses the test's power (to be high enough to deter shirking), and then maximizes the significance level (which minimizes the premium for accidental punishment).

We will concentrate on the simple test that uses only the information of whether total output has exceeded a threshold level T. The principal accepts (H_0: All agents chose high effort) if $Q_{obs} \geq T$ and (H_1: At least one agent chose low effort) if $Q_{obs} < T$. This test is illustrated by Figure 1. The significance level of the test, S, is one minus the probability that the principal mistakenly believes that a single agent chose low effort:

$$S = 1 - \text{Prob}(Q_{n,H} \leq T). \tag{4}$$

The significance level falls as the threshold rises:

$$\frac{dS}{dT} < 0. \tag{5}$$

The power of the test, P, is the probability that a single shirking agent will be detected when present:

$$P = \text{Prob}(Q_{n,L} \leq T). \tag{6}$$

The power rises as the threshold rises:

$$\frac{dP}{dT} > 0. \tag{7}$$

Figure 1: Power and Significance Level

The characteristics of this test are well-known, for it is simply a test for the mean of a normal distribution with known variance against a composite alternative hypothesis. It is, in fact, a textbook example for power and significance level, and it provides the *best critical region* for testing our hypothesis. It is also the *uniformly most powerful* test: not only can it test for the family of alternative hypotheses in which n agents shirk, it is the best such test for all of those alternative hypotheses.[6]

The Compensation Contract.

A natural contract to associate with the threshold test consists of the triplet (T, w, b). Each agent receives w if output exceeds the threshold T and suffers punishment b if output fails to exceed T. We have no reason to believe that the optimal contract lies within this restricted contract space, but as will be explained at the end of Section III, this is not so restrictive as it might seem.

Using the contract (T, w, b), the principal must choose threshold and punishment values such that if the agent shirks, his expected utility is lower than if he works. The expected utility of shirking is based on the probabilities of (a) being caught and rightfully punished (the power), and (b) successful deception and a wage of w. But the agent must compare the expected utility of shirking not with a fixed utility from working hard, but with the expected utility of working hard. The expected utility of working hard is based on the probabilities of (a) being mistakenly punished, and (b) being paid w (the significance level).

The first requirement for a contract is that it deter shirking. Any of a wide range of powers can succeed in this, since the punishment can be chosen appropriately to the power— a bigger punishment for a smaller power. The second requirement for a contract is that it minimize the cost of mistaken punishment. Thus, a high significance level is desirable. But there is a tradeoff between these two requirements: as equations (5) and (7) show, if the contract's threshold increases, the significance level declines while the power rises.

We have assumed that the punishment b reduces the agent's utility without increasing the principal's profit. Since we are interested in comparing the efficiency of different contracts, it is important that the punishment incur deadweight loss in this way rather than just being a transfer. An example of a pure transfer is the forfeiture of a performance bond when both the principal and the agents are risk neutral and collecting the bond incurs no administrative costs. In that case, a contract that results in more frequent bond forfeitures is no less efficient. An agent would only care about expected income, so he would be indifferent between (w = \$190, b = \$10, *50 percent*

probability of punishment) and ($w = \$110$, $b = \$10$, *10 percent probability of punishment*). The expected income is \$100 under either contract.

Punishment creates deadweight loss whenever the punishment's disutility to the agent is greater than its utility to the principal. There are several reasons why punishments that create deadweight loss are commonly used.[7] One reason is that when agents are risk averse even a monetary penalty introduces riskiness into compensation, which hurts the agent without correspondingly benefiting the principal (who, in fact, must also bear risk). A second reason is that any instance of efficient punishment benefits the employer, which raises the problem of deliberate unwarranted punishment. A third reason is that monetary penalties which go beyond a wage of zero to seize part of the agents' assets incur high transactions costs to enforce. A fourth reason is that bankruptcy protection and other legal constraints impose limits on monetary penalties, requiring substitution to nonmonetary penalties such as dismissal or embarassing reprimands. Even if monetary penalties are bounded, the same effect can be achieved by the uses of bonuses. Instead of a high ordinary wage and a severe fine if output is low, the principal pays a low ordinary wage and a large bonus if output is high. But bonuses are particular vulnerable to cheating on the part of the principal; he may misreport that output is low, or even deliberately sabotage output. In addition, the bankruptcy constraint can bite at the level of the principal as well as of the agent. If the principal must pay large bonuses to the entire team, he too can go bankrupt.

Punishments take a number of forms in actual employment. Examples include loss of wage premiums (Becker and Stigler 1974), loss of future wage increases (Doeringer and Piore 1971), damaged reputation (Klein and Leffler 1981), and job search costs after dismissal (Shapiro and Stiglitz 1984). Even if the level of the punishment is exogenous, or the firm gains some advantage from the punishment, the model continues to apply, so long as the gain to the firm is less than the loss to the agent.

In some circumstances it may be possible to avoid the deadweight loss of punishments entirely and attain just as high utility under asymmetric information as under full information. A "boiling-in-oil" contract is a threshold contract under which the principal imposes very severe penalties if output is so low that such an output level would occur with zero probability if every

agent had exerted high effort. Holmstrom (1982) uses such a contract to obtain the first-best outcome in a teams model with hidden effort but no uncertainty in output. Under a boiling-in-oil contract the significance level is equal to one, since the agents exert high effort in equilibrium and the punishment is never inflicted. Such a contract is infeasible here because the support of the normal distribution is the whole range of output. Whatever effort is chosen by the agent, output might be very low, so no threshold can guarantee a significance level of one hundred percent. Holmstrom also shows, using the approach of Mirrlees (1974), that even with uncertainty the first-best outcome can be approximated by a simple threshold contract with large penalties infrequently inflicted. This is the case if there is no bound on penalties and the distribution of output satisfies assumptions ensuring that the product of the penalty and the probability of its infliction can be made vanishingly small.[8]

When boiling-in-oil contracts are infeasible, the principal knows he will sometimes punish agents even when they exert high effort. He also knows that in equilibrium the agents always exert high effort, so every instance of punishment is mistaken. This is a paradox common to every model of costly punishment. In discussing classical hypothesis testing, our language has strayed from that of Bayesian games. If the equilibrium is common knowledge (the standard assumption), the principal should rationally assign zero probability to the presence of a low-effort agent, but we have spoken as if the principal actually believes the test when it indicates low effort and he carries out the punishment. This language is for expositional convenience. In actuality, it is important that both agents and principal have committed to follow the contract and carry out the costly punishment, even though everyone knows that (1) in equilibrium only the innocent are punished and (2) even if an agent did shirk, by the time output is observed it is useless to punish him. Such a situation cannot be avoided, because only by precommitting to carry out punishments can shirking be deterred.

III. The Number of Agents.

We will now try to discover which is better at providing incentive contracts, a large team or a small team. We will compare the significance levels and powers for threshold tests in teams of size n and $n+1$, where the null hypothesis is that all agents are choosing high effort and the alternative hypothesis is that one agent is shirking.

Lemma 1: *If the power of the tests used by teams n and $n+1$ is the same, the significance level is higher for team n.*

Proof: See Appendix.

In view of Lemma 1, as the number of agents increases, the attractiveness of the contract diminishes because of the increase in mistaken detection. The expected wage in equilibrium is always $\overline{q}_h$, since it must yield zero profits, so contracts differ in their attractiveness based on the frequency and size of the punishment in equilibrium. If the significance level of one of the contracts is lower, that contract ends up inflicting a greater expected punishment, and hence is less attractive to agents for a given expected wage. This is true, in particular, if P is the power generated by the optimal contract (T^*, b^*, w^*) for a firm of size n. To increase the size of the team and maintain the same power is costly. This is stated in Proposition 1, which frames the situation in terms of the cost of providing a given level of utility to the agent (in equilibrium, this will be the maximum level that allows an optimizing firm to maintain zero profits).

Proposition 1: *If teams n and $n+1$ choose contracts of the form (T, w, b) to maximize their own profits subject to providing agents with a given level of utility, team n incurs lower costs of contracting.*

Proof: A certain level of power P^* is associated with the contract that maximizes profit for team $n+1$ subject to the reservation utility constraint. Let us denote this optimal contract by (T^*, w^*, b^*). Suppose that team n offers a contract with the same wage w^*, punishment b^*, and power P^* (n's threshold will be different to maintain the same power). By Lemma

1, $S(n, P^*) > S(n+1, P^*)$, so team n will punish the agents less often in equilibrium. Team n's workers will therefore have higher utility than team $(n+1)$'s. This means that the reservation utility constraint is not binding and team n can reduce the wage. This does not depend on whether the agents are risk averse or not. Moreover, team n could also choose the power and punishment optimal for itself, instead of using P^* and b^*, so the wage might be reduced still further. Hence team n has lower costs.
Q.E.D.

It might be appropriate here to repeat the assumptions of the model. Proposition 1 arises from a free-rider problem of sorts, but not a simple free-rider problem. The simple statement that workers shirk more in larger firms because they influence output less is not true, because incentive contracts can sometimes get around the free-rider problem. Holmstrom (1982) shows that if there is no random noise, or particular kinds of random noise, shirking can be prevented, and McAfee and McMillan (1989b) shows that even with general uncertainty, incentive contracts can be found that will deter shirking. The question is how the cost of shirking differs between firms, and that is the question addressed by Proposition 1.

So far, we have maintained the restriction that the contract is of the form (T, w, b). We can easily relax this restriction and allow contracts consisting of a finite number of such triplets, as stated in Proposition 2. Note a caveat absent from Proposition 1: Proposition 2 applies only if the first-best cannot be achieved; otherwise the question of firm size is vacuous.

Proposition 2: *If teams n and $n+1$ choose contracts of the form (T_i, w_i, b_i), $i = 1, \ldots, k$ to maximize their own profits subject to providing agents with a given level of utility, and if these contracts impose real costs, then team n incurs lower costs.*

Proof: Each of the i components of the contract has a particular power P_i and significance level S_i. Lemma 1 says that for each component, firm n can maintain the power at P_i while increasing the significance level. Increasing the significance level is desirable for the reasons discussed in the proof of Proposition 1. (The proof of Proposition 1 did assume that profits equalled

zero, which might not be true of each component, but the proof can easily be adapted to any fixed level of profit.) Firm n is therefore superior in each of the k parts, so agents in firm n can be paid a lower wage than agents in firm $n+1$.
Q.E.D.

Proposition 2 has quite general application because combinations of the (T, w, b) contracts can be used to build step contracts to approximate any continuous contract. Thus, although we have limited ourselves to a class of contracts that might not include the optimal contract, and we can say almost nothing about its form, our results on team size fit a very wide class of contracts.

As an example of how to apply Proposition 2, consider the contract consisting of a flat wage of $\overline{q}_h$ plus a punishment b' inflicted if the team output is less than a particular threshold T'. This contract is outside of the space allowed by Proposition 1, but it can be closely approximated by the two triplets $(T_1 = -\alpha, w_1 = \overline{q}_h, b_1 = 0), (T_2 = T', w_2 = 0, b_1 = b')$, where α is an arbitrarily large number. (This contract is an approximation only because the wage is not quite flat; it falls to zero if output is below $-\alpha$.)

Propositions 1 and 2 continue to be valid even if one requires that contracts use only limited penalties and wages, e.g. $b \in [\underline{b}, \overline{b}]$ and $w \in [\underline{w}, \overline{w}]$, so long as high effort continues to be second-best efficient. This is because Lemma 1 concerns detection, rather than punishment, so the smaller team has a lower level of false punishment for *any* punishment-detection combination, not just the one optimal without the penalty limitation. Hence, a limit on the size of penalties does not remove the advantage of the smaller team.

IV. Identical Effort, Different Abilities

The model so far has been constructed for identical agents who choose effort (moral hazard), but it could also have been constructed for agents whose effort is fixed but who differ in ability (adverse selection).[9] Suppose that agents have either high or low ability, where the proportion of high-ability agents in the economy equals θ, and agents have utility functions $U(w, b)$ that are increasing in the wage w and decreasing in the penalty b. Agents may be either risk averse or risk neutral in the wage. Agent i's output, which depends on his ability and random disturbance, equals

$$q_h = \overline{q}_h + \varepsilon_i \quad \text{or} \quad q_l = \overline{q}_l + \varepsilon_i, \tag{8}$$

where $\overline{q}_h > \overline{q}_l$. These assumptions parallel those in the moral hazard model, but adverse selection requires somewhat more care in modelling because some agents produce high output and some produce low output even in equilibrium, and the equilibrium contract might be either pooling or separating. There are various ways to specify how offers and counteroffers are made in an adverse selection model, and under some specifications the existence of equilibrium is a problem. We do not need to discuss those specifications here; for discussions see Riley (1979) or chapters 8 and 9 of Rasmusen (1989). All that is relevant is whether the equilibrium is pooling or separating.

If a pooling contract were to be part of equilibrium in this game, it would pay the same wage for all outputs and never inflict penalties. For profits to equal zero, the wage would equal the average ability, so the contract would specify a wage of $\theta\overline{q}_h + (1-\theta)\overline{q}_l$ and a penalty of zero. The size of the team would be irrelevant, since no attempt would be made to detect low-ability agents.

If a separating contract were to be part of an equilibrium, the reasoning of Proposition 2 implies that the cost of offering a separating contract to attract just high-ability agents would increase with team size. Any team which offered a contract with a larger team size would have to pay a higher expected wage, and since the smallest teams would earn zero profits under competition with each other, the larger team would earn negative profits. The high-ability agents are thus hired by small teams, and the low-ability agents are hired by teams of any size that use fixed-wage contracts at a wage

of $\overline{q}_l$.

An important difference between the effort and ability versions of the model is that the ability version has implications not only for the size of teams, but also for the distribution of talent among them. Only small teams would be able to offer contracts which attract high-ability agents. Large teams would have to pay higher wages to attract high-ability agents with a contract that still deterred low-ability agents. Hence, high-ability agents would choose small teams that provided contracts which ensured the high quality of working peers. Large teams could still be composed of low-ability agents who are paid a fixed wage, and since size is irrelevant to the efficiency of the fixed-wage contract, some small teams might also be composed of low-ability agents.

V. Empirical Evidence

If we have persuaded the reader that a teams model has something to say about firm size, our model has a number of empirical implications for industrial organization. From an organizational point of view, our results imply that the optimal team size is a single member. A larger team cannot offer as attractive a contract, because it requires a higher probability of mistaken punishment, so if teams can take any size, the optimal team has a single member. This is an implication of any model of managerial diseconomies of scale. But although single-agent firms are common, they certainly do not represent the full range of sizes observed. Indeed, in many industries we observe a wide range of firm sizes at the same time. This diversity of firm sizes is not incompatible with our results, since managerial diseconomies of scale are not the only influence on firm size. If technological economies of scale are present or if external contracting is particularly costly (Coase, 1937; Williamson, 1975), these elements will be traded off against managerial diseconomies of scale to determine the optimal firm size. In addition, our model makes no prediction for the size of firms that employ low-effort or low-ability agents on fixed-wage contracts. Such firms can be either large or small. Finally, firms need not rely solely on the measurement of team output to detect shirking or low ability. They also have the option of monitoring effort or testing ability. If firms can detect shirking or low ability for a fixed cost per agent, then as firm size increases and the cost of incentive contracts rise, testing or monitoring become cheaper than incentive contracts. If there are economies of scale to monitoring and testing, then large firms using those methods might coexist with small firms using incentive contracts.

Our model suggests that although large firms and small firms can exist in the same industry, they will differ in their management styles and employment contracts. Large firms will offer fixed-wage contracts and make heavier use of monitoring, testing, and easily observable employee characteristics to control productivity. Small firms will link pay and performance more closely, extract greater effort, and hire the most talented employees (conditioning on observable variables). Small firms may also be willing to employ those low-quality individuals rejected by the screening of large firms, because the small firms can pay them an appropriately low wage using output-based contracts.

We will compare these empirical implications with evidence from earlier studies and data from the 1979 Current Population Survey. The CPS is designed to be representative of the entire U.S. labor force. Employees were surveyed, and they estimated the size of their firm by choosing between five size categories: 1) 1-24 employees, 2) 25-99 employees, 3) 100-499 employees, 4) 500-999 employees, and 5) 1000 or more employees. We use wage regressions to examine the predicted differences between large and small firms in employment contracts. We use regressions of self-reported hours worked to examine the predicted differences in effort.

A. Use of Observable Employee Characteristics in Wage-Determination

A first prediction is that large firms will rely more heavily than small firms on directly observable worker characteristics such as education or seniority, as a substitute for performance-based contracts. Small firms, which are more efficient at detecting low effort and ability, will more closely link pay and performance. The implications of this can be looked at in two ways: (1) individual variables such as tenure will explain wages better for large firms, and (2) the set of such variables will explain wages better for large firms.

Table 1 presents descriptive information for individuals in each of the five size categories. Table 2 presents five equations (for the five firm sizes) using the CPS data to regress the log of hourly wages on tenure with the firm, work experience, education, and various dummy control variables such as industry, occupation, location, and union membership.[10]

Table 1: DESCRIPTIVE STATISTICS

Table 2: LOG WAGE EQUATIONS: THE EFFECT OF TENURE

Tenure has a significant effect on compensation in all size categories. Consistent with our hypothesis, the effect of tenure is greater in large firms than in small for nearly the entire range of tenure values (up to 41 years). The estimated coefficients suggest, for example, that an employee with two years tenure receives 1.30% in additional income for remaining an additional

year at a small firm, but 1.84% at a large firm.[11] At the entire sample's mean tenure of 8.25 years, an additional year yields 1.26% additional compensation in a small firm, but 1.59% additional compensation in a large firm (1000+ employees).[12]

There are many specific reasons why tenure might matter more in large firms. Large firms might have more complex bureaucracies, or rely more heavily on deferred compensation, or attach more importance to learning about ability and effort over time. But all of these specifics are subheadings of the general reason we propose: that compensation in large firms relies less on current output and more on other factors than does compensation in small firms.

The other regression variables we expected to be important were experience and education. The coefficients for "Other Experience" are significant, but roughly one third the size of tenure's, and without important variation across firm size, except for relative unimportance in the largest category. Education also shows no clear size-related pattern. To the extent that previous experience and education are collinear with ability, we would not expect much variation in these parameters across firm sizes, since these attributes are easily observed.

A second way to interpret the wage equations is to look at how well wages are explained by the right-hand-side variables in aggregate. The model predicts greater residual variance in the small-firm regression, because small firms link pay to performance instead of to the right-hand-side variables. The easiest way to check for residual variance is to look at the R^2 values in Table 2. The R^2 for small firms is .358, whereas the values for the four larger categories are .402, .447, .450, and .421. These results are generally consistent with our prediction, although the small value for the largest firms is anomalous.

We can also test for the difference in explanatory power more formally. Under our model the wage equations are misspecified for small firms, since performance is a relevant and omitted variable. But we can take as our null hypothesis that the wage equations are correctly specified, that large firms are identical to small ones in their use of the explanatory variables, and that

the random disturbances follow the same normal distribution for all firms. Under this null hypothesis, the variance of the residual is identical for the five categories, and the ratio of the squared standard errors from any two of the five regressions follows the F-distribution, with degrees of freedom equalling the sample size for each regression. Table 2 shows the F-statistics for the differences between the standard errors of the smallest firms and each of the four other categories. The regression for the smallest firms has a significantly greater standard error than for any of the other size categories, rejecting the null of no difference.[13]

B. Self-Reported Hours of Work.

A second prediction is that as a result of these size-related differences in employment contracts, effort will be lower in large firms than in small firms. CPS respondents reported the number of hours they worked during the week preceding the survey, and we can use their self-reported hours as an indication of effort. Hours worked is clearly not a perfect measure of overall effort, since hours worked measures only the duration of effort and not its intensity. Indeed, if hours worked were a perfectly accurate measure of effort for all employees, then all employees would presumably be paid by the hour. But our tolerance for measurement error can be considerably greater than the tolerance of managers, since managers, unlike researchers, must directly compensate for the uncertainty imposed by errors in measurement. Hence, for our purposes, a substantial correlation between the duration of effort and overall effort, as seems reasonable, is sufficient. The likelihood of such a correlation between true effort and hours worked is partly contingent on managers not using hours worked as a measure of effort. If all employees in the large firm were paid by the hour, the correlation between hours and effort would decline as employees adjusted their behavior toward longer, but less intensive effort. The measurement of hours worked by government statisticians does not have this same behavior-altering effect on employees.

Table 3 presents separate regressions for hourly and non-hourly employees of hours worked per week on firm size, work experience, education, and various dummies including industry, union membership, occupation, and location. Only the coefficients for the dummy size categories are displayed,

with very large size (1000+ employees) as the excluded category. The results indicate that full-time, non-hourly workers employed in very small firms (1-24 employees) on average work 2.4 hours per week more than non-hourly workers employed in very large firms. Those employed in small firms (25-99 employees) on average work 1.4 hours per week more than non-hourly workers in very large firms. Note also that the relationship between firm size and hours worked appears to be non-linear since hours worked do not differ significantly among those employees in the three large firm size categories.

Table 3: FIRM SIZE AND HOURS WORKED

Our model suggests that since time is a directly monitored input for hourly employees, they should should work the same number of hours in large and small firms. Table 3 shows that even for hourly employees there is a significant negative relationship between size and hours, but not as strong or as significant as for non-hourly employees. Hourly employees of very small firms work .85 hours more per week than hourly employees of very large firms. Hourly employees in small and medium-sized firms (100-499 employees) work on average just over .5 hour more per week than in very large firms. Reasons for these size-related differences must be found outside our model, but the results help to calibrate the extent to which the coefficient on firm size in the non-hourly regressions is due to omitted variables, and show it to be small.

Cross-industry empirical work of this kind, while useful for finding whether an effect is widespread, is also subject to the criticism that it might be driven by omitted industry variables, however many control variables are included in the regressions. Another approach is to examine firms within a single industry. An example is Zenger (1989), which compares contracts at large and small firms using survey responses from a sample of engineers who had left two large high-technology firms. The findings suggest that contracts at smaller firms involve greater equity ownership, link firm performance more closely to pay, involve less formal monitoring, and impose greater employment risk. Moreover, among engineers departing the two firms, those of higher ability left for smaller firms and those of lower ability left for larger firms. Finally, the engineers in small firms worked more hours per week, consistent with the CPS regressions of Table 3.

C. Previous Work on Firm Size and Employment Contracts

Various investigators have found a relationship between firm size and the employment contract. Garen (1985) examines wage models from the National Longitudinal Survey and finds a marginally significant relationship between wages and the interaction of ability (measured by test scores) and firm size (measured by the percentage of the industry's labor force employed in firms with more than 500 employees). Bishop (1987) similarly finds that productivity has an important positive effect on wages in small, non-union establishments, but almost no effect in large unionized establishments. Medoff and Abraham (1980) examine the compensation practices of two large firms and confirm a weak link between pay and performance. These results support the conclusion that ability and compensation are more closely associated in small firms than large firms.

Our model may be particularly applicable to R&D settings, where individual outputs are difficult to discern and teamwork is essential. Also consistent with our reasoning and these results is the common, although not undisputed empirical finding that R&D is more efficiently performed by small and medium-sized than by large firms (see Chapter 3 of Kamien and Schwartz, 1982). Our model predicts that large firms cannot efficiently offer contracts that induce high effort and attract high ability. The survey data from Zenger (1989) supports this view.

Another prediction is that earnings at small firms should vary more than at large firms. High-ability employees should all be attracted to small firms, while low-ability employees might work at large or small firms. It is well known that average earnings are higher in large firms than in small firms, contrary to our model's prediction, though it is not clear why this is so, since the effect persists even after controlling for observable indicators of worker quality. In fact, the thorough study of Brown and Medoff (1989) finds that the effect is just as strong for *piece-rate* workers at large firms. This may be the result of large firms employing testing or monitoring procedures that weed out workers who are low-quality in terms of either observables or unobservables. Then a more complete prediction of our model would be that small firms will include some firms with incentive contracts employing high-output workers and some firms with fixed-wage contracts employing

very-low-ability workers who are rejected by the large firms. In addition to our findings, Brown and Medoff, Garen (1985), and Stigler (1962) have found greater variability in earnings among employees of small firms than among employees of large firms.

Our results linking firm size and effort are also consistent with experimental studies in psychology examining the effects of group size. These studies confirm negative relationships between group size (ranging from 2 to 8 members) and individual effort in rope-pulling, brainstorming, hand-clapping, shouting, and use of an air pump.[14] They also find a curvilinear relationship consistent with the regression results of Table 3: the marginal effect of the Nth person on the efforts of group members is less than the marginal effect of the (N-1)th individual.

VI. Concluding Remarks

This paper develops a model of the relationship between team size and the efficiency with which contracts based on team output can resolve problems of moral hazard and adverse selection. The model implies that contracts based on team output are more efficient in identifying and deterring low effort and low ability for small teams than for large teams. We argue that these conclusions are also relevant to firm size. Consequently, large firms are more likely than small firms to avoid contracts that base workers' compensation on firm output. Instead, large firms offer fixed-wage contracts that do not closely link compensation to ability or effort, but use seniority or other observable criteria to determine compensation. In addition, large firms will aggressively test for low ability. Small firms will identify and reward ability and effort by linking pay and firm output. As a consequence of these contractual differences, we predict that small firms will induce higher effort and employ low-ability workers rejected by large firms as well as high-ability workers attracted by incentive contracts. Empirical results are consistent with these predictions.

Our model and empirical analysis provide a partial explanation for the managerial or organizational diseconomies of scale assumed in price theory and transactions-cost economics. The costs of organizing rise with firm size because larger firms are less efficient than smaller firms in offering contracts that induce high effort and attract high-ability workers.

Figure 1: Power and Significance Level

Table 1: DESCRIPTIVE STATISTICS

Table 2: LOG WAGE EQUATIONS: THE EFFECT OF TENURE

Table 3: MEAN HOURS WORKED PER WEEK BY FIRM SIZE

APPENDIX: Proof of Lemma 1

If just one agent chooses low effort, the outputs for teams n and $n+1$ are

$$Q_{n,L} = \overline{q}_l + \varepsilon_1 + \sum_{i=2}^{n}(\overline{q}_h + \varepsilon_i) \tag{9}$$

and

$$Q_{n+1,L} = \overline{q}_l + \varepsilon_1 + \sum_{i=2}^{n+1}(\overline{q}_h + \varepsilon_i). \tag{10}$$

The two variables $Q_{n,L}$ and $Q_{n+1,L}$ both have normal distributions, because they are the sums of normally distributed random variables. (This is true whether the random variables are independent or not.) Their expected values are

$$\mu_{n,L} = \overline{q}_l + (n-1)\overline{q}_h \tag{11}$$

and

$$\mu_{n+1,L} = \overline{q}_l + n\overline{q}_h. \tag{12}$$

The variance of output depends on the team size and the correlation between the errors, but not on whether agents shirk. If the errors are independent, then

$$\sigma_n^2 = n\sigma^2 \tag{13}$$

and

$$\sigma_{n+1}^2 = (n+1)\sigma^2. \tag{14}$$

If the errors are perfectly correlated, then

$$\sigma_n^2 = n^2\sigma^2 \tag{15}$$

and

$$\sigma_{n+1}^2 = (n+1)^2\sigma^2. \tag{16}$$

In either case, or for any positive degree of correlation between errors (or even a sufficiently small negative correlation),

$$\sigma_{n+1} > \sigma_n. \tag{17}$$

If the power equals P for either size team, then

$$P = Prob(Q_{n,L} \leq T_n) = Prob(Q_{n+1,L} \leq T_{n+1}). \tag{18}$$

Using normality,

$$P = Prob(Q_{n,L} \leq T_n) = \Phi\left(\frac{T_n - \mu_{n,L}}{\sigma_n}\right) = \Phi\left(\frac{T_n - \overline{q}_l - (n-1)\overline{q}_h}{\sigma_n}\right) \tag{19}$$

and

$$P = Prob(Q_{n+1,L} \leq T_{n+1}) = \Phi\left(\frac{T_{n+1} - \mu_{n+1,L}}{\sigma_{n+1}}\right) = \Phi\left(\frac{T_{n+1} - \overline{q}_l - n\overline{q}_h}{\sigma_{n+1}}\right). \tag{20}$$

Let us define

$$A_1 \equiv \frac{T_n - \overline{q}_l - (n-1)\overline{q}_h}{\sigma_n} \tag{21}$$

and

$$A_2 \equiv \frac{T_{n+1} - \overline{q}_l - n\overline{q}_h}{\sigma_{n+1}}. \tag{22}$$

From the fact that (19) and (20) equal the same P, we can conclude that $A_1 = A_2$.

If all the agents choose high effort, the outputs are

$$Q_{n,H} = \sum_{i=1}^{n} (\overline{q}_h + \varepsilon_i) \tag{23}$$

and

$$Q_{n+1,H} = \sum_{i=1}^{n+1} (\overline{q}_h + \varepsilon_i). \tag{24}$$

These two variables also have normal distributions. The significance levels for the given power are

$$\begin{aligned} S(n,P) &= Prob(Q_{n,H} \geq T_n) \\ &= 1 - Prob(Q_{n,H} \leq T_n) \end{aligned} \tag{25}$$

and

$$\begin{aligned} S(n+1,P) &= Prob(Q_{n+1,H} \geq T_{n+1}) \\ &= 1 - Prob(Q_{n+1,H} \leq T_{n+1}). \end{aligned} \tag{26}$$

These two significance levels are not necessarily equal. We can rewrite them using the normality assumption and the definitions of A_1 and A_2. Equation (25) becomes

$$\begin{aligned} S(n,P) &= 1-\Phi\left(\tfrac{T_n-\mu_{n,H}}{\sigma_n}\right) \\ &= 1-\Phi\left(\tfrac{T_n-n\overline{q}_h}{\sigma_n}\right) \\ &= 1-\Phi\left(\tfrac{T_n-\overline{q}_l-(n-1)\overline{q}_h-\overline{q}_h+\overline{q}_l}{\sigma_n}\right) \\ &= 1-\Phi(A_1-\tfrac{(\overline{q}_h-\overline{q}_l)}{\sigma_n}). \end{aligned} \tag{27}$$

In the same way, equation (26) becomes

$$\begin{aligned} S(n+1,P) &= 1-\Phi\left(\tfrac{T_{n+1}-\mu_{n+1,H}}{\sigma_{n+1}}\right) \\ &= 1-\Phi\left(\tfrac{T_{n+1}-(n+1)\overline{q}_h}{\sigma_{n+1}}\right) \\ &= 1-\Phi(\tfrac{T_{n+1}-\overline{q}_l-n\overline{q}_h-\overline{q}_h+\overline{q}_l}{\sigma_{n+1}}) \\ &= 1-\Phi(A_2-\tfrac{(\overline{q}_h-\overline{q}_l)}{\sigma_{n+1}}). \end{aligned} \tag{28}$$

By equation (17), $\sigma_{n+1} > \sigma_n$, so

$$\frac{(\overline{q}_h-\overline{q}_l)}{\sigma_n} > \frac{(\overline{q}_h-\overline{q}_l)}{\sigma_{n+1}}. \tag{29}$$

It follows from (29) and the fact that $A_1 = A_2$, that

$$\Phi\left(A_1-\frac{(\overline{q}_h-\overline{q}_l)}{\sigma_n}\right) < \Phi\left(A_2-\frac{(\overline{q}_h-\overline{q}_l)}{\sigma_{n+1}}\right), \tag{30}$$

so by equations (27) and (28) it is true that $S(n,P) > S(n+1,P)$.
Q.E.D.

REFERENCES.

Albanese, Robert and David Van Fleet. 1985. "Rational Behavior in Groups: The Free-Riding Tendency." 10 *Academy of Management Review* 244-255.

Baker, George, Michael Jensen, and Kevin J. Murphy. 1988. "Compensation and Incentives: Practice vs. Theory." 43 *Journal of Finance* 593-616.

Becker, Gary and George Stigler. 1974. "Law Enforcement, Malfeasance, and the Compensation of Enforcers." 3 *Journal of Legal Studies* 1-18.

Bendor, Jonathan and Dilip Mookherjee. 1987. "Institutional Structure and the Logic of Ongoing Collective Action." 81 *American Political Science Review* 129-154.

Bickel, Peter and Kjell Doksum. 1977. *Mathematical Statistics.* San Francisco: Holden-Day.

Bishop, John. 1987. "The Recognition and Reward of Employee Performance." 5 *Journal of Labor Economics* S36-S56.

Boyd, Robert and Peter Richerson. 1988. "The Evolution of Reciprocity in Sizable Groups." 132 *Journal of Theoretical Biology* 337-356.

Brown, Charles and James Medoff. 1989. "The Employer-Size Wage Effect." 97 *Journal of Political Economy* 1027-1059.

Calvo, Guillermo, and Wellisz, Stanislaw. 1980. "Technology, Entrepreneurs, and Firm Size." 95 *Quarterly Journal of Economics* 663-678.

Coase, Ronald. 1937. "The Nature of the Firm. 4 *Economica* 386-405.

Coase, Ronald. 1988. "The Nature of the Firm: Influence." *Journal of Law, Economics, and Organization* 33-47.

Doeringer, Peter, and Piore, Michael. 1971. *Internal Labor Markets and Manpower Analysis.* Boston: D.S. Heath and Company.

Farrell, Joseph and Eric Lander. 1989. "Competition Between and Within Teams: The Lifeboat Principle." *Economics Letters*, forthcoming.

Frank, Robert. 1985. *Choosing the Right Pond.* Oxford: Oxford University Press.

Garen, John. 1985. "Worker Heterogeneity, Job Screening, and Firm Size." 93 *Journal of Political Economy* 715-739.

Holmstrom, Bengt. 1982. "Moral Hazard in Teams." 13 *Bell Journal of Economics* 324-340.

Holmstrom, Bengt. 1988. "Agency Costs and Innovation." Yale University Working Paper.

Kaldor, Nicholas. 1934. "The Equilibrium of the Firm." 44 *Economic Journal* 60-76.

Kamien, Morton and Nancy Schwartz. 1982. *Market Structure and Innovation.* Cambridge: Cambridge University Press.

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

Klein, Benjamin and Keith Leffler. 1982. "The Role of Market Forces in Assuring Contractual Performance." 89 *Journal of Political Economy* 615-641.

Latane, Bibb. 1984. "The Psychology of Social Impact." 36 *American Psychologist* 343-356.

Lucas, Robert. 1978. "On the Size Distribution of Business Firms." 9 *Bell Journal of Economics* 508-523.

Maddala. 1977. *Econometrics.* New York: McGraw Hill, 1977.

Marshall, Alfred. 1920. *Principles of Economics.* 8th edition. Reprinted, London: MacMillan Press, 1977.

McAfee, R. Preston and McMillan. 1989a. "Organizational Diseconomies of Scale." Mimeo, IRPS, University of California, San Diego, 10 April 1989.

McAfee, R. Preston & John McMillan. 1989b. "Optimal Contracts for Teams." Mimeo, IRPS, University of California, San Diego, undated.

Medoff, J., and K. Abraham. 1980. "Experience, Performance, and Earnings." 95 *Quarterly Journal of Economics* 703-736.

Milgrom, P., and J. Roberts. 1988a. "An Economic Approach to Influence Activities in Organizations." 94 Supplement *American Journal of Sociology* S154-S179.

Milgrom, P., and J. Roberts. 1988b. "Bargaining Costs, Influence Costs, and the Organization of Economic Activity." Stanford University Department of Economics Working Paper.

Mirrlees, James. 1974. "Notes on Welfare Economics, Information and Uncertainty." In Balch, McFadden, and Wu, eds., *Essays on Economic Behavior under Uncertainty.* Amsterdam: North Holland.

Porter, Robert. 1983. "Optimal Cartel Trigger Price Strategies." 29 *Journal of Economic Theory* 313-338.

Rasmusen, Eric. 1989. *Games and Information.* Oxford: Basil Blackwell, 1989.

Riley, John. 1979. "Informational Equilibrium." 47 *Econometrica* 331-359.

Scherer, Frederick, Alan Beckenstein, Erich Kaufer and R. Dennis Murphy. 1975. *The Economics of Multi-Plant Operation.* Cambridge, Mass: Harvard University Press.

Schumpeter, Joseph. 1950. *Capitalism, Socialism and Democracy.* 3rd Edition. New York: Harper and Row, 1975.

Shapiro, Carl and Joseph Stiglitz. 1984. "Equilibrium Unemployment as a Worker Discipline Device." 74 *American Economic Review* 433-444.

Shavell, Steven. 1985. "Criminal Law and the Optimal Use of Nonmonetary Sanctions as a Deterrence." 85 *Columbia Law Review* 1232-1263.

Sraffa, Piero. 1926. "The Laws of Returns Under Competitive Conditions." 36 *Economic Journal* 535-50.

Stigler, George. 1958. "The Economies of Scale."1 *Journal of Law and Economics* 54-71.

Stigler, George. 1962. "Information in the Labor Market." 70 (Supplement) *Journal of Political Economy* 94-105.

Stigler, George. 1964. "A Theory of Oligopoly." 72 *Journal of Political Economy* 44-61.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation & Welfare*, Oxford: Basil Blackwell, 1986.

Viner, Jacob. 1931. "Cost Curves and Supply Curves." 3 *Zeitschrift fur Nationalokonomie* 23-46. Reprinted in *Readings in Price Theory*, ed. George Stigler and Kenneth Boulding. Homewood, Illinois: Irwin, 1952.

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

Williamson, Oliver. 1985. *The Economic Institutions of Capitalism.* New York: The Free Press, 1985.

Zenger, Todd. 1989. "Organizational Diseconomies of Scale: Pooling vs. Separating Labor Contracts in Silicon Valley." Ph.D. dissertation, University of California, Los Angeles.

Footnotes.

We would like to thank Steven Lippman, Ivan Png, Emmanuel Petrakis, Steven Postrel, Robert Topel, and Sang Tran for helpful comments. The data was made available in part by the Inter-university Consortium for Political and Social Research.

1. See page 265 of Marshall (1920): "In other words, we say broadly that while the part which nature plays in production shows a tendency to diminishing return, the part which man plays shows a tendency to increasing return. The *law of increasing return* may be worded thus:— An increase of labour and capital leads generally to improved organization, which increase the efficiency of the work of labour and capital."

2. Schumpeter (1950), p. 101. More recent work along these lines includes Lucas (1978) and Calvo and Wellisz (1980), who argue that high-ability managers will go to large firms where their greater capacity to manage can be put to better use. This is a valid point, but our model will assume that the individual contribution of an agent to the team's output is the same regardless of the team's size. We will try to isolate just one effect, and, like technological economies of scale, the increasing sphere for talent could swamp the disincentive effect we find for large teams.

3. See, e.g., Stigler's 1958 article on the Survivor Principle.

4. See Williamson (1985, Chapter 6) for a more complete discussion of the constraints faced by large firms in replicating small firm incentives through multiple subunits.

5. Strictly speaking, the probability of avoiding a Type I error is the *size* of the test, and a test of size 0.95 is a test of *significance level* 0.95, 0.94, 0.93, and so forth. Since the term "size" is somewhat obscure (it is omitted from the index of many statistics texts), we will use "significance level" here, with the understanding that we mean the test's highest significance level.

6. One textbook that uses this test as an example in discussing these statistical points is Bickel & Docksum (1977). See Chapters 5 and 6, and especially pages 168-71, 192, and 198.

7. The question of why costly punishments are used is also a lively question in the economics of crime. Shavell (1985) is a recent reference.

8. Our output distribution satisfies those assumptions, which means that any size team can achieve "almost" the first-best if penalties are unbounded.

9. A third possibility is that both ability *and* effort vary between agents. We do not address that here; for a discussion, see McAfee and McMillan (1989b).

10. Some of these control variables may depend on whether it is efficient for the firm to use incentive contracts. Since unions frown on incentive pay, for example, it might be that large firms, for which a flat wage might be more efficient, would resist unionization less strongly. If that is true, then by controlling for unionization we underestimate the effect of firm size.

11. These values were determined by calculating the effect of tenure at 3 years and then subtracting the effect of tenure at 2 years. For instance, the value 1.84% was calculated: $3(.0194) + 3^2(-.002) - 2(.0194) + 2^2(-.0002)$.

12. We have also performed the regressions for a subsample of CPS data covering just professional and technical employees, whose output is particularly hard to measure. The results are similar. Similar regressions were also performed for a subsample that included only non-union males. These results were also similar and, indeed, stronger than the results in Table 2.

13. F(120,120) = 1.35 at the 5 percent level, 1.53 at the 1 percent level (Maddala 1977, pp. 510-11). The lowest F-statistic in Table 2 is 1.494, and the lowest sample size is 755, so every test is clearly significant.

14. For surveys of this literature, see Albanese and Fleet (1985) and Latane (1981).

SCRAPS.

Lemma 1's statement that a larger sample produces a worse test is counterintuitive at first sight, because we are used to thinking about statistical tests for the difference between the mean and a constant. For a fixed significance level, such as 95 percent, the power of such a test increases with the sample size if the errors are independent, which seems to imply the opposite of Lemma 1. But that is not the relevant test here. Rather, we are testing for the presence of one shirking agent. The value in our alternative hypothesis is not a constant, but a variable that gets closer to the value in the null hypothesis as the sample size increases. Speaking

loosely, Lemma 1 says that the increasing difficulty of distinguishing the alternative hypothesis outweighs the increasing precision of the estimate.

Another curious feature of Lemma 1 is that shirking in the larger team is harder to catch despite the fact that as the team size increases, the variance of output per agent may decrease. If the disturbances are independent, the variance of output per agent equals

$$v_n^2 = \frac{\sigma^2}{n}, \tag{31}$$

and if the errors are perfectly correlated it equals

$$v_n^2 = \sigma^2. \tag{32}$$

Hence, one might think that team size should not matter if the disturbances are perfectly correlated, and that the larger team should be superior if the disturbances are less than perfectly correlated. This is wrong. Output per team member is not the relevant variable; its variance may fall to zero as n increases, but that does not tell us that detecting a single deviator becomes easier, since the individual disturbance's share of the total noise also falls as n increases.

From: tzengerDate: Mon, 4 Dec 89 15:56:33 pst Message-Id: ¡8912042356.AA25681
To: fac, $as Subject : revised document$

Dear Eric,

This is the revised document. I made the change we talked about on page 4, but I think it may need to be changed further in conjunction with the revisions you make on page 19 regarding the Holmstrom article. I have moved the proof to the Appendix and have dropped the counterintuitiveness discussion. I have also altered H1 and the su surrounding discussion.

Todd

Pfeffer, Jeffrey and Langton, Nancy. 1988. "Wage Inequality and the Organization of Work: The Case of Academic Departments." 33 *Administrative Science Quarterly* 588-606.

———————————————————————