Low-Powered Incentives

LOW-POWERED INCENTIVES

May 1999

Abstract

We study low-powered incentives in a model that captures important features of workplaces in which incentive-pay approaches are minimally relevant. Our motivation is that incentive pay, while not rare, is clearly far less common than are agency problems; many firms with agency problems nonetheless pay fixed compensation and offer continued employment to all but those workers judged "unsatisfactory" according to largely subjective criteria. We find that low-powered incentives can achieve efficient outcomes in simple workplaces and function surprisingly well even when the environment is characterized by unobservable performance heterogeneity and a high degree of complementarity among workers.

KEYWORDS: incentives, o-ring technology, subjective performance assessments

JEL CLASSIFICATIONS: J33

Joseph A. Ritter
Research Department
Federal Reserve Bank of St. Louis
411 Locust Street
St. Louis, MO 63102

Lowell J. Taylor
H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213

Economists' natural inclination in studying employment relationships is to design incentives. Thus economists have generated a large literature on incentive pay (recent surveys are Gibbons (1998) and Prendergast (1999)). Yet incentive pay, while not rare, is clearly far less pervasive than are agency problems; many firms with agency problems nonetheless pay fixed compensation and offer continued employment to all but clearly unsatisfactory employees. In this paper we study the structure of an employment relationship when measures of performance on which explicit or implicit contracts can be based are either absent or very costly, or where attempts to exploit them result in highly dysfunctional outcomes. Thus the firm is limited to "low-powered" incentives: fixed current compensation coupled with inability to reliably distinguish satisfactory from unsatisfactory workers.

Much of the previous research in this area has pointed to theoretical or empirical failures of incentive-pay schemes of various sorts, but little attention has been devoted to understanding the strengths and weaknesses of low-powered incentives *per se.* In particular, various authors have highlighted the ease with which information imperfections can disrupt the proper functioning of incentive-pay schemes. Holmstrom and Milgrom's (1994) conclusion captures the flavor of this line of work: "The use of low-powered incentives within the firm, while sometimes lamented as one of the major disadvantages of internal organization, is also an important vehicle for inspiring cooperation and coordination." Empirically, the incidence of incentive pay is certainly not particularly high, and its magnitude is typically small relative to fixed compensation. MacLeod and Parent (1997) document that the compensation of a large fraction of U.S. workers does not include any sort of incentive pay. Bewley (forthcoming) found that, although most managers would like to use incentive pay, many believe that, for various reasons, they cannot.

Two distinctions are important for understanding the role of low-powered in-

centives. The first is between objective and subjective assessments of performance.[1] In principle, current or future pay can be conditioned on the measures of performance chosen by the firm, but the choice of subjective measures imposes additional constraints. In particular, the nonverifiability of subjective assessments narrows the set of possible equilibria. On the other hand, the strength of subjective assessments is that a manager can reasonably determine whether a worker is doing a "good" job in, say, a multi-tasking environment or where "output" is intangible. Even if explicit performance measurements were available for each task, the combinatorial demands of optimally balancing them in an explicit contract quickly become overwhelming (MacLeod and Parent, 1999).

The second distinction is the extent of feedback between performance assessments and compensation; this is what defines "low-powered" incentives. Although, arguably, virtually every employer connects subjective assessments with pay in some way, the evidence suggests that employers do not differentiate among employees as sharply as economists might expect. Salary and rating compression is widespread, for example. There are good reasons for employers' reluctance to tie pay closely to subjective assessments: In addition to the problems that plague the use of explicit performance measurement, influence activities and human biases can easily dampen or destroy incentives based on subjective assessments.

In this paper, therefore, we study the design of low-powered incentives for an information structure that captures important features of workplaces in which incentive-pay approaches are minimally relevant: Either contractible variables appropriate for incentive-pay do not exist, or they result in dysfunctional compensation schemes that should be avoided. The model we study allows the firm to use

---

[1] Subjective assessments have been formally modeled as common knowledge measurements that cannot be verified by a third party (Baker, 1992; Baker, Gibbons, and Murphy, 1994). A further distinction can be drawn between common knowledge and private assessments.

only one type of information: managers' impressions or opinions, formed in the course of performing other duties. We assume these impressions are a costless, but private and noisy measure of workers' performance. Since our goal is to understand the functioning of low-powered incentives, we initially restrict the use of this information to determining whether to retain a worker. Thus there is no feedback between the assessments and current or future compensation.

Simpler models of low-powered incentives have been studied in the literature on deferred compensation and efficiency wages, but these models have been meant primarily to illustrate features of labor market equilibrium, rather than to illuminate the economics of information flows inside the firm (for example, Lazear, 1979; Shapiro and Stiglitz, 1984). Thus, these authors make a number of simplifying assumptions that obscure the operation of low-powered incentives: "Effort" is taken to be a discrete variable with only one economically viable value. The information structure of the firm is crude and unrealistic; "monitoring" is an entirely exogenous, costless, and common-knowledge process. The interaction of "effort" and "monitoring" with the production technology is left in the background.[2]

The theory we develop is illuminating on several dimensions. First, our basic formulation with no worker heterogeneity allows us to delineate clearly the role that performance assessments play in the mechanism of low-powered incentives. Two striking findings emerge: (1) Despite a rarified information environment, low-powered incentives achieve efficient outcomes. (2) Only the split of the joint surplus is affected by the quality of performance assessments. Under broad assumptions, bad managers—those who are unbiased but unable to evaluate performance accurately— or difficult information environments are more costly to the firm, but do not otherwise distort employment relationships. Compensation is higher, but

[2] A notable recent exception is Mehta (1998), whose model addresses the relationships among technology, monitoring, and span of control.

turnover and performance are not affected.

The introduction of unobservable heterogeneity generates our most interesting insights. In the spirit of the multi-tasking literature, it seems plausible that workers often know better than their supervisors how to do their jobs on a day-to-day basis. Following Baker, Gibbons, and Murphy (1994), we represent this idea by variation in the employee's productivity that is observable only to the employee. Low-powered employment policies are robust in that they generate outcomes that are close to optimal when there are small amounts of unobservable heterogeneity.

Most important, with heterogeneity in performance, compensation policies are not independent of the production technology. Employment policies interact strongly with the degree of complementarity among workers, and this is one of the areas in which low-powered employment policies seem to excel. In particular, when the production technology is characterized by strong complementarity among workers' performance levels (dubbed "o-ring" technology by Michael Kremer (1993)), low-powered incentives respond appropriately by inducing close-to-efficient performance from low-productivity workers. As complementarity among performance levels increases, firms find poor performance at the low end more and more undesirable. For most parameter values in our model, firms control this downward variation in performance by raising compensation and, somewhat paradoxically, by *reducing* the probability of dismissal for low-productivity workers.

One interpretation of our low-powered incentives is that after the efficient application of the various feasible high-powered incentive mechanisms, most firms are likely to face a residual set of agency issues that can be addressed only through the use of these low-powered incentives. For instance, limitations on the size of "prizes" in a rank-order tournament may mean that they cannot completely resolve agency problems the firm faces. The importance of this residual pool of agency troubles will vary from firm to firm, depending on the nature of the production technology

and information flows in the firm. To understand cross-sectional variation, then, it is necessary to understand how the high- and low-powered incentives interact with one another and with the technology of production. A simple rank-order tournament serves as a vehicle to illustrate the principles involved in this interaction. In the basic model, implementation of higher-powered incentives shifts surplus back to the firm without interfering with efficiency. In the model with heterogeneity, matters are more interesting. Higher-powered incentives are relatively ineffective in boosting performance of the least productive workers, the most important group when there is significant complementarity among workers.

We turn now to the layout of the basic model, which we study in detail in Section II, deferring consideration of unobservable heterogeneity to Section III. In Section IV we study the interaction of low- and high-powered incentives (in the form of a rank-order tournament).

I. The Model

To focus on the functioning of low-powered incentives, we abstract from life-cycle effects by assuming that workers are infinitely-lived (or, equivalently, face a Poisson probability of death). Later we introduce a simple rank-order tournament as a surrogate for the effects of promotion ladders and other intertemporal arrangements based on subjective assessments. In each period workers choose a scalar performance level p, known only to the worker.

The performance levels of N workers determine the output of a profit-maximizing firm:

$$Y = G(p^1, p^2, \ldots, p^N)$$

with $\partial G/\partial p^i > 0$. We assume that the firm cannot be profitable if all workers supply $p = 0$: $G(0, \ldots, 0) = 0$. When G is additively separable in p^i the production function is similar to those used in models of team production such as those formu-

lated by Bengt Holmstrom (1982).[3] Activities such as fruit picking by a team of agricultural workers would fit this characterization. We are especially interested in cases where the technology makes workers' performance levels highly complementary. Extreme degrees of complementarity generate "o-ring" technology like that studied by Michael Kremer (1993). A symphony orchestra, for example, fits this description: No matter how well the other 47 members of an orchestra play, if the principal second violin plays out of tune, the Mahler is ruined.[4]

Total output Y is arguably contractible, but, since we are interested in environments where low-powered incentives prevail, we assume that Y provides no usable information about p_i. We also rule out strategic interactions among workers: Workers do not condition their behavior on the behavior of other workers. These two assumptions make our information structure quite dissimilar to the team production problem. Formally, our assumption about the information content of Y means that we simply do not consider compensation schemes that condition pay on Y. Informally, we interpret this assumption to mean that either there are no variables related to a worker's contribution to firm value on which either an explicit or implicit contract can be based, or that the firm should not use them because they would provide dysfunctional incentives for any of the various reasons discussed in the literature.[5]

Since workers appear identical to the firm, they are treated alike, so we work

---

[3] In team production models, a central focus is the moral hazard from the opportunity to free-ride. As will be apparent shortly, we do not address this issue, as compensation in our model is not conditioned on total output.

[4] Kremer (1993) gives other examples. Kremer and Maskin (1996) discuss the potential importance of o-ring production for understanding current labor market trends.

[5] Prendergast (1999) notes: "Perhaps the most striking aspect of observed contracts is that the Informativeness Principle, i.e., that all factors correlated with performance should be included in a compensation contract, seems to be *violated* in many occupations."

with the function

$$g(p^i) = G(p^i, p^{(-i)}),$$

where $p^{(-i)}$ denotes the performance levels of workers other than i. We assume that $g(p)$ has continuous first and second derivatives and is concave. The concavity of $g(p)$ is best interpreted in the present context as reflecting complementarity among workers, but plays no significant role in the model until Section III, where we discuss it in depth.[6]

Initially, we assume that performance p changes one-for-one with the worker's disutility from work. Interpreting p as an effort level causes no confusion until Section III, though we prefer the term "performance," which hints at the range of problems, such as multi-tasking, that might induce the firm to adopt a low-powered strategy. A worker's utility in each period is the wage less her performance level: $w - p$. In Section III we introduce random variation in the disutility ("effort") required to achieve a given performance level.

Employed workers supply one unit of labor in the period. They have a discount factor $\beta < 1$.[7] The expected value of a worker's alternative activities is V^a. Hiring and termination are costless to the firm.

The firm's information about the performance level of a particular worker is minimal: The firm receives a noisy signal (or summary statistic) x of the worker's performance,

$$x = p + \epsilon.$$

We assume that ϵ has zero mean, and that its distribution is single-peaked with density f and distribution function F. We interpret x as a manager's impression or

---

[6] This is just the most interesting interpretation; the technology could be $G(p^1, \ldots, p^N) = g(p^1) + \ldots + g(p^N)$, which displays no interdependence among performance levels, even if g is concave.

[7] A constant exogenous probability of separation would enter the model in a fashion almost identical to β.

opinion of how well the worker is performing. The realization of x is indisputably private information and, therefore, cannot form the explicit basis for any compensation policy, or is at least suspect in this regard. Rank-order tournaments with a fixed pool of "prizes" are the least implausible kind of high-powered scheme in this environment, and we later examine in some detail how they might function in our low-information environment. Although x is private information, the distribution F summarizes characteristics of the manager and production environment that are common knowledge. A clueless manager has a relatively diffuse distribution for ϵ. We assume the manager behaves in the firm's interests, though, of course, this is a subject of considerable independent interest.

Although we focus on an interpretation that ascribes F to the manager, the nature of firms' production processes also generates cross-sectional variation in F. If workers are physically separated from managers, for example, managers' impressions would probably be more diffuse. The span of a manager's control, explicitly modeled by Mehta (1998), would also affect F.

Our information structure differs in an important way from others that use subjective performance evaluation. In particular, Baker, Gibbons and Murphy (1994) use a subjective assessment that is the firm's evaluation of a worker's contribution to the firm (y in their paper). In that paper, the subjective assessment is not contractible, but *is* common knowledge between the firm and worker. Therefore, it is possible for the worker to condition a strategy on whether the firm pays an "implicit-contract bonus" in line with the firm's subjective assessment; the worker can easily assess whether the firm has violated the implicit contract. In the present context, however, the subjective assessment x is *not* known to the worker, which effectively makes it impossible to use x to adjust compensation in any useful way without an extraordinary level of trust. The existence of such trust is likely to be rare, however, so we examine only equilibria which do not exploit it. Note that if

compensation were a continuous function of x, the worker would need to be concerned with both large-scale, relatively easy to detect, cheating and minor, difficult to detect, chiselling.

II. Employment Policy with Homogenous Performance

Since workers do not interact strategically with one another, the model is a repeated game between the firm and a single worker with the following order of play in each round: (1) The firm offers a wage w (which is contractible). (2) The worker responds with a performance level $p \geq 0$. (3) Nature plays x using the distribution F. (4) The firm pays w. (5) The firm decides whether to retain the worker or end the game. We focus on repeated play of the unique Bayesian Nash equilibrium of this stage game in which the worker is retained if and only if x exceeds an endogenous threshold. We assume that this threshold $\bar{x}$ is common knowledge—workers observe the frequency of terminations.

A. The Employee's Problem

Let $\hat{p}$ be a worker's best response to an employment policy $\psi = \{w, \bar{x}\}$ that satisfies the worker's participation constraint: $V(\hat{p}; \psi) \geq V^a$, where $V(\hat{p}; \psi)$ is the maximum utility that can be achieved by accepting ψ. The lifetime utility of an employee who chooses performance level p today and reverts to $\hat{p}$ tomorrow is given by

$$V(p; \psi) = w - p + \beta\big[F(\bar{x} - p)V^a + \big(1 - F(\bar{x} - p)\big)V(\hat{p}; \psi)\big]. \tag{1}$$

The employee maximizes $V(p; \psi)$ by choosing $p \geq 0$. The worker's best response is quite well behaved:

Proposition 1: *An employment policy $\psi = \{w, \bar{x}\}$ that satisfies the participation constraint implies a unique best response $\hat{p}$. Further, $\hat{p}$ exceeds the termination threshold $\bar{x}$.*

Proof: First note that $V(p;\psi)$ is continuous in p, and $\lim_{p\to\infty} V(p;\psi) = -\infty$, so $V(p;\psi)$ must have a maximum for $p > 0$.[8] Thus for today's best action to be $\hat{p}$, we must have $V'(\hat{p};\psi) = 0$, which reduces to

$$\beta f(\bar{x} - \hat{p})[V(\hat{p};\psi) - V^a] = 1. \tag{2}$$

We also have

$$V''(p;\psi) = \beta \frac{f'(\bar{x} - p)[V^a - V(\hat{p};\psi)]}{1 - \beta\big((1 - F(\bar{x} - \hat{p})\big)},$$

which is negative long as $f'(\bar{x} - p) > 0$. Thus $V(p;\psi)$ is concave where $\bar{x} - p < 0$. Therefore, a solution $\hat{p}$ to (2) is a maximum if $\bar{x} - \hat{p} < 0$. The maximum is unique, since two maxima with $\bar{x} - \hat{p} < 0$ would bracket a minimum. Since $f(\epsilon)$ is single-peaked, this is impossible . ∎

According to Proposition 1, workers are terminated only when their perceived performance is strictly less than their actual performance. The worker, recognizing the imperfection of the signal, establishes a buffer between her actual performance, $\hat{p}$, and $\bar{x}$, the perceived performance level that results in termination. Note that this is a claim about the behavior of the worker in response to any sufficiently desirable ψ, not only an equilibrium ψ. Intuition suggests that the buffer increases as the signal becomes more diffuse. We will show in Proposition 3 that under quite general assumptions about the distribution of x, that intuition is borne out.

An outsider able to observe both $\hat{p}$ and x might misinterpret outcomes in two ways. The observer might conclude that managers tolerate substandard performance, firing workers only when they see performance well below the norm. Alternatively, the observer might interpret $\hat{p} - \bar{x}$ as a "gift" of performance exceeding some standard or required level, but, in fact there is no reciprocity in this model.

---

[8] An arbitrary ψ might imply the corner solution $p = 0$. But since $G(0, 0, \ldots, 0) = 0$, no operating firm would offer a ψ that induced this corner solution, so we ignore this possibility without significant loss of generality.

Evaluating (1) at $p = \hat{p}$ and solving for $V(\hat{p}; \psi)$, then substituting into (2) produces

$$w = \hat{p} + (1 - \beta)V^a + \frac{1 - \beta\big(1 - F(\bar{x} - \hat{p})\big)}{\beta f(\bar{x} - \hat{p})}, \tag{3}$$

which implicitly defines the worker's best response $\hat{p}(\psi)$. Since $w - \hat{p}$ is the flow of utility from the job and $(1-\beta)V^a$ is the flow value of alternative activities, equation (3) guarantees that any ψ for which $\hat{p} > 0$ also satisfies the participation constraint.

It is interesting to note that $\hat{p}$ is continuous in ψ. In other words, if $\{w, \bar{x}\}$ is optimal for the firm, a small decrease in w will not result in a discrete jump to $p = 0$, as it does in models that assume 0/1 effort decisions.

B. Profit Maximization

The firm's objective is to maximize profits, which it must do subject to the constraint imposed by the worker's best response function $\hat{p}(\psi)$:

$$\max_{\psi} g(\hat{p}(\psi)) - w.$$

This problem turns out to be much less cumbersome if we observe that equation (3) has a dual interpretation as specifying the minimum w required to induce performance $\hat{p}$ for a given $\bar{x}$.[9] We thus write (3) as $w = w(\hat{p}, \bar{x})$ and use it to formulate an equivalent profit maximization:

$$\max_{\bar{x}, \hat{p}} g(\hat{p}) - w(\hat{p}, \bar{x}). \tag{4}$$

We will continue to use $\hat{p}$ to denote the worker's best response, while using p^* to denote the performance induced by the firm's optimal choice of ψ. An interesting result follows easily from the modified profit-maximization problem (4):

---

[9] That is,

$$w(\hat{p}, \bar{x}) = \min_{w}\{w \text{ subject to } (3)\}.$$

Proposition 2: *The performance level p^* induced by the optimal employment policy ψ^* in a firm with a homogeneous workforce is first-best:*

$$g'(p^*) = 1.$$

Proof: The first-order conditions for (4) are

$$0 = g'(p^*) + \frac{\partial w(p^*, \bar{x}^*)}{\partial p^*}$$
$$0 = \frac{\partial w(p^*, \bar{x}^*)}{\partial \bar{x}}.$$

Since both $\hat{p}$ and $\bar{x}$ are parameters in the minimization implied by the dual interpretation of (3), we can apply the envelope theorem to $w(\hat{p}, \bar{x})$ (differentiate the right-hand side of (3)). By inspection,

$$\frac{\partial w(\hat{p}, \bar{x})}{\partial \hat{p}} = 1 - \frac{\partial w(\hat{p}, \bar{x})}{\partial \bar{x}},$$

and thus $g'(p^*) = 1$. ∎

The efficient performance level achieved by low-powered incentives justifies part of our claim that bad managers do not distort employment relationships when performance is homogeneous. For a wide class of densities $f(\cdot)$, including normal, logistic, and nonstandard t variables, as well as mixtures of them, we can characterize the equilibrium employment policy in detail.

Proposition 3: *Suppose that the density f depends on a parameter σ, with σ^2 linearly related to the variance, and that f can be written*

$$f(\epsilon; \sigma) = \frac{1}{\sigma} h\left(\frac{\epsilon}{\sigma}\right)$$

where h is differentiable and does not depend on σ except via ϵ/σ. Then

i. The retention rate does not depend on σ. That is, the firm chooses w and $\bar{x}$ so that $(\bar{x}^* - p^*)/\sigma = b^*$, where b^* does not depend on σ.

ii. Compensation is linearly increasing in σ:

$$w(p^*, \bar{x}^*) = p^* + (1 - \beta)V^a + \frac{1 - \beta\Gamma(b^*)}{\beta h(b^*)}\sigma. \tag{5}$$

Proof: Make the change of variables $b = y/\sigma$, so that

$$f(\sigma b) = \frac{h(b)}{\sigma}.$$

It is easy to show that

$$f'(\sigma b) = \frac{h'(b)}{\sigma^2}$$

From profit maximization we have

$$0 = \frac{\partial w(p^*, \bar{x}^*)}{\partial \bar{x}} = 1 - \frac{[1 - \beta(1 - F(\bar{x}^* - p^*))]f'(\bar{x}^* - p^*)}{\beta f(\bar{x}^* - p^*)^2}$$

Solving for the retention probability and substituting for f and f' makes it clear that the retention probability is independent of σ:

$$\Gamma(b^*) \equiv 1 - F(\bar{x}^* - p^*) = \frac{1}{\beta} + \frac{h(b^*)^2}{h'(b^*)}$$

Substitution into (3) produces equation (5). ∎

Combined with Proposition 2, this result demonstrates that bad managers or difficult environments make it costly to achieve efficient performance. One might expect that in an increasingly noisy environment the firm would respond by trading off lower expected performance for lower compensation. Instead, increasing σ only moves the worker farther away from her participation constraint (while $\sigma = 0$ makes the participation constraint bind). The mechanism works in the following way: Higher σ reduces the value of the job to the worker because the probability of termination is higher. The parties respond by increasing the spread between p and $\bar{x}$ (returning the probability of termination to its original level) and by increasing the wage (directly increasing $V(p; \psi)$). This is, however, the only distortion that results; the performance level not affected. As in the standard principal-agent problem, where a linear contract usually has a negative intercept, some sort of lump-sum transfer is needed if the firm is to force the worker onto her participation constraint.

III. Low-Powered Incentives with Unobservable Heterogeneity

We turn next to the most important contribution of our paper—the study of low-powered incentives in a firm confronted with unobservable worker heterogeneity. Though we do not model it here, there is one obvious reason why the our low-powered incentive structure, which relies on firing workers with particularly low observed performance, will be effective in dealing with heterogeneity: the turnover acts as a device to filter permanent, unobservable heterogeneity. Low ability workers, for example, will make systematically different choices or get systematically different outcomes for the same choices, thus changing their probability of surviving the $\bar{x}$ screen. Explicit modeling of that phenomenon would require careful attention to the process by which the firm learns about a given worker through a series of draws on x.

In this section we focus instead on the implications of a more subtle form of unobservable heterogeneity—transitory variation in the work environment. If workers are better positioned than supervisors to observe these stochastic elements in the production process, then the motivational effect of any incentive scheme that relies on the supervisor's assessment will vary unobservably with circumstances. As we demonstrate shortly, this in turn implies that employment policies interact in important and systematic ways with the nature of the production technology itself.

A. A Simple Model of Heterogeneity

Following Baker, Gibbons, and Murphy (1994), suppose that the effort required to achieve a given level of performance is random, so that the worker's performance is $p = e/\eta$, where e is effort—the disutility incurred in supplying performance p. The random variable η is binary with $\eta_1 > \eta_2$, $\Pr\{\eta = \eta_1\} = \theta$, and $\mathrm{E}\{\eta\} = 1$. We will see that higher performance is associated with lower marginal disutility η_2. Each realization of η is independent of past realizations and those of other

workers. We assume that η is known only to the worker and thus not contractible. We continue to assume that $x = p + \epsilon$ and output is $g(p)$. However, we now have

$$g(p_j^i) = \mathrm{E}\{G(p_j^i, p^{(-i)})\}, \tag{6}$$

where the expectation is taken with respect to the (binomial) distribution of other workers' productivity realizations, that is, $\eta^{(-i)}$. As in Section II, the problem we solve should be understood as designing the employment policy ψ that should be offered to worker i holding other workers' performance at the levels, p_1^* and p_2^*, induced by the optimal policy ψ^*.

We first consider the worker's best response to employment conditions ψ. The expected lifetime utility of a worker faced with η_j is

$$V(p_j, \eta_j, \psi) = w - \eta_j p_j + q\beta V^a + \beta[F(\bar{x} - p_j)V^a + \big(1 - F(\bar{x} - p_j)\big)\mathrm{E}_\eta\{V\}, \tag{7}$$

where $\mathrm{E}_\eta\{V\} = \theta V(\hat{p}_1, \eta_1, \psi) + (1 - \theta)V(\hat{p}_2, \eta_2, \psi)$. The first-order condition for maximizing V with respect to p_j is

$$\eta_j = \beta[\mathrm{E}_\eta\{V\} - V^a]f(\bar{x} - \hat{p}_j). \tag{8}$$

The structure of the employee's maximization is much the same as in the simpler model, and thus Proposition 1 applies to both $\hat{p}_1$ and $\hat{p}_2$. Taking expectations of both sides of (7) with respect to η gives an expression for $\mathrm{E}_\eta\{V(\hat{p}, \eta; \psi)\}$ that can be used in (8). Rearrangement produces

$$\begin{aligned} w = {} & \theta\eta_1\hat{p}_1 + (1 - \theta)\eta_2\hat{p}_2 + (1 - \beta)V^a \\ & + \frac{1 - \beta[1 - \theta F(\bar{x} - \hat{p}_1) - (1 - \theta)F(\bar{x} - \hat{p}_2)]}{\beta}\left(\frac{\eta_j}{f(\bar{x} - \hat{p}_j)}\right) \end{aligned} \tag{9}$$

for $j = 1, 2$. Equations (9) implicitly define the worker's best responses $\hat{p}_1$ and $\hat{p}_2$ to employment policy ψ. Note that equations (9) imply

$$\frac{f(\bar{x} - \hat{p}_1)}{f(\bar{x} - \hat{p}_2)} = \frac{\eta_1}{\eta_2}. \tag{10}$$

We digress briefly to state the characteristics of the best-response functions, which offer some intuition for the pattern of numerical results we observe and discuss below.

Proposition 4: *For all ψ that satisfy the worker's participation constraint, the worker's best-response functions obey:*

$$1 = \frac{\partial \hat{p}_1}{\partial w} + \frac{\partial \hat{p}_1}{\partial \bar{x}}, \qquad 1 = \frac{\partial \hat{p}_2}{\partial w} + \frac{\partial \hat{p}_2}{\partial \bar{x}}.$$

Suppose $f(\bar{x} - p_2)/f(\bar{x} - p_1)$ is increasing in $\bar{x}$ (that is, f possesses the monotone likelihood ratio property), then at the profit-maximizing employment policy, ψ^*,[10]

$$\frac{\partial \hat{p}_1}{\partial w} > 1, \qquad \frac{\partial \hat{p}_1}{\partial \bar{x}} < 0, \qquad \frac{\partial \hat{p}_2}{\partial w} > 0, \qquad \frac{\partial \hat{p}_2}{\partial \bar{x}} > 0.$$

Sketch of Proof: The proof is long, so we outline it only. The first part comes from straightforward differentiation and manipulation of equations (9). There are three steps in proving the second part. First, profits are $\theta g\big(\hat{p}_1(\bar{x}, w)\big) + (1-\theta) g\big(\hat{p}_2(\bar{x}, w)\big) - w$. The first-order condition for $\bar{x}$ implies that $\partial \hat{p}_1/\partial \bar{x}$ and $\partial \hat{p}_2/\partial \bar{x}$ have opposite signs. Second, the first part of the proposition implies that either $\partial \hat{p}_1/\partial w$ or $\partial \hat{p}_2/\partial w$ is greater than 1, while the other is less than one. Third, differentiating (10) with respect to w gives a relationship between these two derivatives, and the monotone likelihood ratio property is sufficient to guarantee $\partial \hat{p}_1/\partial w > \partial \hat{p}_2/\partial w$. ∎

The central ideas of Proposition 4 are sensible. The first two equalities establish that the firm can always increase performance by some specific amount, say Δp_j by simply effecting identical corresponding increases in the wage and dismissal threshold, $\Delta w = \Delta \bar{x} = \Delta p_j$. The change in w compensates the worker for the

[10] The monotone likelihood ratio assumption is frequently invoked in agency theory. Distributions in this class have the property that higher signals are always relatively more likely to have come from the distribution with the higher mean. For details, see Milgrom (1981). Many of the distributions that satisfy the hypotheses of Proposition 3 also possess the monotone likelihood ratio property, for example, the normal and logistic distributions. However, not all such distributions display the monotone likelihood ratio property over all possible values.

additional disutility incurred, and the change in $\bar{x}$ exactly restores the probability of termination to its old value. The derivatives of p_j with respect to the wage are both, not surprisingly, positive. The derivatives with respect to $\bar{x}$ are more interesting. With homogenous performance, the firm, for any given w, adjusts $\bar{x}$ so as to maximize performance: $\partial p/\partial \bar{x} = 0$. Here the firm chooses a compromise, $\bar{x}$, which is "too high" from the perspective of achieving the highest possible p_1 and "too low" from the perspective of achieving maximum p_2.

B. Low-Powered Incentives Are Robust

The firm chooses ψ to maximize $\theta g(\hat{p}_1) + (1-\theta)g(\hat{p}_2) - w$ subject to equations (9). Reasoning similar to that leading to Proposition 2, gives us a rather interesting generalization of Proposition 2:

Proposition 5:

$$\theta g'(p_1^*) + (1-\theta)g'(p_2^*) = \theta\eta_1 + (1-\theta)\eta_2 = 1. \tag{11}$$

The proof can be found in an appendix. Note that, combined with (10), Proposition 5 implies that $p_1^* < 1 < p_2^*$. Proposition 5 looks innocuous, but in fact demonstrates the robustness of low-powered incentives on three levels. First, equation (11) shows that Proposition 2 is robust to infrequent heterogeneity. If θ is low so that $\eta = \eta_1$ is a rare event, the low-powered incentive mechanism pushes incentives toward efficiency when the common event occurs; if θ is small, (11) cannot hold unless $g'(p_2^*) \approx \eta_2$.

Second, low-powered incentives will be robust to small amounts of heterogeneity ($\eta_1/\eta_2 \approx 1$). Since $f(\cdot)$ is continuous and $\bar{x} - p_2^*$ and $\bar{x} - p_1^*$ are to the left of the single peak of $f(\cdot)$, equation (10) implies that $p_1^* \approx p_2^*$. Combined with (11) this means that $p_1^* \approx p_2^* \approx 1$; small amounts of unobservable heterogeneity do not result in grossly dysfunctional behavior.

Third, and most interesting, suppose we normalize performance so that $g(1) = g'(1) = 1$ and then make g progressively more concave while keeping $g(1) = g'(1) = 1$.[11] (In the limit $g(\cdot)$ has a corner at 1.) In the present context, increasing concavity of $g(\cdot)$ is most naturally interpreted as increasing complementarity among workers' performance levels; more concavity moves $g(p)$ closer to o-ring technology. To illustrate, suppose that $G(\cdot)$ is a constant-elasticity-of-substitution function. Ignoring the expectation in (6) (which is just the sum of functions with the this shape) and, for simplicity, assuming that all other workers work at performance level $\bar{p} = 1$, we have $g(p) = G(p, \bar{p}, \ldots, \bar{p}) = A[p^a + (n-1)\bar{p}^{\,a}]^{\frac{1}{a}}$. The concavity of $g(p)$ can be increased in the way specified above by moving the elasticity parameter a toward $-\infty$, while changing A to keep the scale from changing at $p = \bar{p} = 1$.

As g becomes more concave in this sense, the efficient level of p_1 converges to 1. Proposition 5 implies that the actual level of p_1 also converges to 1. If it did not, $\theta g'(p_1)$ would exceed 1. In other words, low-powered incentives respond appropriately to the gross inefficiency that would come from allowing performance to drop too far as $G(\cdot)$ approaches an "o-ring" technology.

In terms of concavity, Proposition 5 gives only a limiting result. A stronger property can be proven, if we are more precise about the kind of deformation of g we have in mind. Let $\psi^* = \{\bar{x}^*, w^*\}$ be the optimal employment policy when the production function is $g(p)$. We will say that a production function $\widetilde{g}(p)$ is "closer to o-ring" or "more concave around 1" than $g(p)$ if four conditions hold:

$$(i) \quad \widetilde{g}(1) = \widetilde{g}'(1) = g(1) = g'(1) = 1,$$

$$(ii) \quad \widetilde{g}'(p) > g'(p) \text{ for } p < 1,$$

$$(iii) \quad \widetilde{g}'(p) \leq g'(p) \text{ for } p > 1,$$

[11] We hold scale fixed in order to isolate the effects of concavity. The exact point at which we do so is somewhat arbitrary, though 1 is a natural choice because $\mathrm{E}\{\eta\} = 1$ and $p_1^* < 1 < p_2^*$.

$$(iv) \quad \theta\widetilde{g}'(p_1) + (1-\theta)\widetilde{g}'(p_2) > 1. \tag{12}$$

Inequality (12) says that changing the production function to $\widetilde{g}$ without adjusting ψ causes (11) to be violated in a particular direction. Our definition is complicated, but intuitive: What we have in mind is simply that the firm becomes more averse to downward variation in performance, while upward variation confers less advantage. That is, the production function bends down more for $p < 1$ than for $p > 1$.[12] Figure 1 illustrates the effects of increasing concavity in this way.

The intuition that Proposition 5 suggests about the robustness of low-powered incentives when there is significant complementarity among workers is borne out in our next result:

Proposition 6: *Suppose that $\widetilde{g}(p)$ closer to o-ring than $g(p)$. Then $\widetilde{p}_1 > p_1$.*

The proof can be found in the appendix. Proposition 6 demonstrates that a firm whose technology is closer to o-ring responds appropriately to that fact when using low-powered incentives. Together, Propositions 4 and 6 allow us to understand how variations in worker complementarity across firms can matter for the structure of low-powered incentives. As a thought experiment, consider a firm that wishes to increase its initial optimal p_1^*, while at the same time becomes less concerned about p_2^*. (This would a firm's response, for example, to a marginal shift toward o-ring technology.) Proposition 4 tells us that one way to achieve the desired increase in p_1 would be to simply increase w and $\bar{x}$ by an equal amount. However, since $\bar{x}$ is always "too high" from the perspective of optimizing p_1, a more clever solution would use a smaller increase in w, accompanied by a decrease (or perhaps a relatively small increase) in $\bar{x}$.

---

[12] The last piece of our definition, inequality (12), is partly endogenous in the sense that "increasing concavity" depends on where you start. Avoiding this requires a more restrictive definition, for example changing $g(p)$ only for $p < 1$. Also, θ enters because very low θ will make the effect of increasing the marginal product of η_1 workers unimportant.

In fact, this is the pattern we observe in numerical experiments with the model: Changing the technology toward o-ring, usually increases the equilibrium value of w and decreases that of $\bar{x}$, though when θ is high $\bar{x}$ increases. With o-ring-like technology, it is critical that the performance of the lowest-performing agent be sufficiently high. Low-powered incentives generate this outcome by using relatively high wages, but, surprisingly, these are typically paired with relatively low dismissal thresholds.

We have been unable to rule out analytically the possibility that both w and $\bar{x}$ decline under the increasing complementarity scenario, but in numerical experiments we have been unable to find this outcome. (Proposition 6 rules out the possibility that w is lower and $\bar{x}$ higher.) Thus the results shown in Figure 1 appear to be representative. The low-powered incentive mechanism responds to greater complementarity among workers by inducing better performance from low-productivity workers, simultaneously over-engineering incentives for high-productivity workers.

C. Other Specifications of Unobservable Heterogeneity

Brief consideration of two variations on our assumption about where the unobservable heterogeneity appears in the model is informative. Suppose first that η affects output, but not utility, so that output is $\eta g(p)$ and flow utility is $w - p$. Again, η is known only to the worker. (Recall our assumption that an individual worker's output cannot be isolated, so the firm has no usable information, contractible or otherwise, on $\eta g(p)$.) We continue to assume $x = p + \epsilon$. This model is relatively easy to analyze. The worker's best-response function is identical to the model with homogeneous workers because g does not have any influence on the worker's incentives. Thus $\hat{p}_1 = \hat{p}_2$; although workers see high and low marginal product days, they do not respond to this information.

Our second alternative model is identical to the first, except that we assume the

firm has information about output, rather than effort: $x = \eta g(p) + \epsilon$. (It is almost impossible to interpret this assumption in the context of a multi-worker firm unless $G(p^1, \ldots, p^N)$ is additively separable in $p^1, \ldots, p^N$, but the analysis is somewhat informative, nonetheless.) The signal x is still only a manager's impression and not contractible. Since η is not observed by the firm, it cannot set state-dependent values of $\bar{x}$. Since utility is linear in effort, the worker will choose effort levels to make the probability of termination constant across states. Thus $\eta_1 g(\hat{p}_1) = \eta_2 g(\hat{p}_2)$.

In the first alternative model, low-powered incentives cannot induce the worker to vary p to take advantage of high-productivity states. When the firm monitors output, however, the situation is far worse. Effort is *negatively* correlated with productivity. A high value of η shifts the mean of x upward, making the probability of termination too low from the worker's point of view, relative to the constant marginal utility of reduced effort.

IV. Rank-Order Tournaments

The information available to agents in our model makes the use of most types of incentive pay implausible. As we have mentioned, since x is entirely subjective and not observable to the worker, a compensation scheme based directly on x would be easily manipulated by the employer. However, Malcomson (1984) and others have observed that rank-order tournaments with preannounced prizes, may have a critical advantage when the employer is tempted by this form of moral hazard. If the employer can credibly commit to awarding a fixed amount of prize money, and workers have some assurance that influence activities will not distort rankings,[13] they may be willing to respond favorably to a rank-order tournament. In this

[13] Prendergast and Topel (1996) have argued that influence activity discourages the use of incentive pay. Also, Lazear (1989) has noted that when cooperation among workers is important, salary compression (that is, movement toward lower-powered incentives) is optimal.

sense a tournament is the incentive-pay scheme that is closest to our low-powered incentives.

In this section we introduce a tournament similar to those described by Lazear and Rosen (1981). Our objective is to study how this type of high-powered incentive interacts with low-powered incentives in our low-information environment.

We begin by reinstating the assumption that workers supply homogeneous performance. Suppose there are just two workers. The tournament pays wage w_H to the worker with the higher realization of x, and $w_L < w_H$ to the other worker. We assume that the firm can commit to paying out $w_H + w_L$, so average compensation is known to be $\overline{w} = (w_H + w_L)/2$. The firm continues to terminate workers when x falls below a threshold $\bar{x}$, so $\psi = \{w_H, w_L, \bar{x}\}$. Finally, we assume a symmetric Nash equilibrium of the game between the two workers, so they supply the same p.

Letting superscript a or b denote the worker, the value to worker a of supplying performance level p today and the optimal level $\hat{p}$ tomorrow is

$$V(p;\psi) = w_H \Pr\{x^a > x^b\} + w_L[1 - \Pr\{x^a > x^b\}] - p \\ + \beta\big[F(\bar{x} - p)V^a + \big(1 - F(\bar{x} - p)\big)V(\hat{p};\psi)\big].$$

$\Pr\{x^a > x^b\}$ can be written in a more useful form:

$$\begin{aligned}\Pr\{x^a > x^b\} &= \Pr\{p^a + \epsilon^a > p^b + \epsilon^b\} = \Pr\{\epsilon^a > p^b + \epsilon^b - p^a\} \\ &= \int_{-\infty}^{\infty} \Pr\{\epsilon^a > p^b + \epsilon^b - p^a | \epsilon^b\}\, f(\epsilon^b)\, d\epsilon^b \\ &= \int_{-\infty}^{\infty} [1 - F(p^b - p^a + \epsilon^b)] f(\epsilon^b)\, d\epsilon^b\end{aligned}$$

The worker's best response must satisfy $V'(\hat{p};\psi) = 0$. Since $\hat{p}_2 = \hat{p}_1$ in a symmetric equilibrium,

$$\frac{\partial \Pr\{x^a > x^b\}}{\partial p^a} = \int_{-\infty}^{\infty} f(\epsilon^b)^2\, d\epsilon^b,$$

and the first-order condition can be written as

$$(w_H - w_L)\int_{-\infty}^{\infty} f(\epsilon^b)^2\, d\epsilon^b + \beta f(\bar{x} - \hat{p})\big[V(\hat{p};\psi) - V^a\big] = 1.$$

Substituting for $V(\hat{p};\psi)$ gives

$$\bar{w} = \hat{p} + (1-\beta)V^a + \left[1 - (w_H - w_L)\int_{-\infty}^{\infty} f(\epsilon^b)^2\, d\epsilon^b\right] \frac{1-\beta\big(1-F(\bar{x}-\hat{p})\big)}{\beta f(\bar{x}-\hat{p})}, \quad (13)$$

which, like (3), implicitly defines the best response $\hat{p}(\psi)$. In the absence of the "prize," $(w_H - w_L)$, equation (13) reduces to (3). The mechanics of profit maximization are much the same as with low-powered incentives. In particular, $g'(p^*) = 1$. Not surprisingly, then, if the tournament can be implemented, it does not interfere with efficiency.

If $f(\cdot)$ is of the form assumed in Proposition 3, the fraction on the right-hand side of (13) does not depend on σ and we have

$$\bar{w} = \hat{p} + (1-\beta)V^a + \frac{1-\beta\Gamma(b^*)}{\beta h(b^*)}\left[\sigma - (w_H - w_L)\int_{-\infty}^{\infty} h(b)^2\, db\right]. \quad (14)$$

The first thing to note about (14) is that, in principle, the tournament does work in this environment; a "prize" of $w_H - w_L$ helps lower average compensation. Indeed, as in Lazear and Rosen (1981), if $w_H - w_L$ can be made large enough, the worker can be pushed onto her participation constraint (that is, $\bar{w} = \hat{p} + (1-\beta)V^a$). In fact, though the specifics of our model are quite different than Lazear and Rosen's, assuming normality and setting the bracketed term in (14) to zero (pushing the worker onto her participation constraint) gives $w_H - w_L = 2\sqrt{\pi}\sigma$, which is equivalent to Lazear and Rosen's result under normality.[14]

The second thing to note about (14) is that to the extent there is a residual agency problem, low-powered incentives work on top of the high-powered incentives

[14] Equation (14) shows that the size of the tournament "prize" $(w_H - w_L)$ needed to move the worker to her participation constraint, depends not only on the variance of the manager's signal, but also on higher moments through the term $\int h(b)^2 db$. For example, if x has a normal distribution, the prize needs to be 7.1 percent higher than with a logistic distribution having the same variance. Note that the logistic distribution has higher kurtosis than the normal distribution, with more variation appearing in the form of signals far away from the mean.

in exactly the same way we have outlined in previous sections; only the level of compensation changes. Neither the performance level nor the turnover rate is affected by the presence of the tournament. The equilibrium of the model is exactly the same as it would have been using low-powered incentives alone but with a variance equal to the bracketed term in (14).

It is easy to see exactly how well known constraints on tournaments work in our low-information environment. A bad manager (or, more accurately, workers' perception that the manager is bad), or a difficult environment, may mean that σ is so high that it is impossible to achieve a large enough gap between w_H and w_L. Alternatively, influence activities may limit the feasible size of the prize.

The advantages of low-powered incentives that we demonstrated in Section III are reinforced by considering the interaction of unobserved heterogeneity with the tournament. Recall that when there is significant complementarity among workers' performance levels, optimal low-powered incentives discourage downward variation in performance. It turns out that rank-order tournaments are likely to be ineffective, and possibly counterproductive, when faced with this kind heterogeneity. Intuitively, tournaments are ineffective in an o-ring environment for the following reason. A worker who gets a bad draw (η_1) realizes that at least one other worker is likely to have gotten a good draw, so she is at a big disadvantage in the tournament. Therefore the prize provides little, if any, motivation. Why expend extra effort to win the tournament when you know the deck is already stacked against you? With o-ring production, the firm is most interested in motivating the workers with bad draws. Yet the tournament provides the strongest motivation to workers with good draws—workers for whom the marginal value of effort is near zero.

To illustrate this mechanism more precisely, consider the problem from the perspective of the worker a. Once she draws η_j, the value of supplying performance

p_j today and the optimal levels $\hat{p}_1$ and $\hat{p}_2$ in the future is

$$V(p_j;\psi) = w_H \Pr\{x^a > x^b\} + w_L[1 - \Pr\{x^a > x^b\}] - \eta_j p_j \\ + \beta\big[F(\bar{x} - p_j)V^a + \big(1 - F(\bar{x} - p_j)\big)\mathrm{E}_\eta\{V(\hat{p}, \eta; \psi)\}\big],$$

The first-order conditions for maximizing V are:

$$\begin{aligned} f(\bar{x} - \hat{p}_1)\beta\big[\mathrm{E}_\eta\{V\} - V^a\big] + \frac{\partial \Pr\{x^a > x^b\}}{\partial p_1}[w_H - w_L] &= \eta_1 \\ f(\bar{x} - \hat{p}_2)\beta\big[\mathrm{E}_\eta\{V\} - V^a\big] + \frac{\partial \Pr\{x^a > x^b\}}{\partial p_2}[w_H - w_L] &= \eta_2. \end{aligned} \tag{15}$$

Note that the bracketed terms in (15) do not depend on the realization of η. The first term on the left side of each equation is the effect of marginal effort on retaining a valuable job times the value of the job. The second term is the effect of marginal effort on the probability of winning the tournament times the prize. The sum of these two terms must equal the disutility of marginal effort.

Now consider the effects of introducing a small tournament, without changing either average compensation or $\bar{x}$. Holding $\hat{p}_1$ and $\hat{p}_2$ fixed, the value of the job is unaffected by this tournament. If $\theta < 1/2$, it is possible to show that

$$\frac{\partial \Pr\{x^a > x^b\}}{\partial p_1} < \frac{\partial \Pr\{x^a > x^b\}}{\partial p_2}.$$

In other words, worker a believes the marginal effect of effort on winning is relatively small when she herself has drawn η_1 if it is unlikely that worker b has also drawn η_1. That is, she does not expect effort to significantly enhance her prospects of winning if she is unlikely to be playing in a fair tournament. In a setting with more than two workers, the probability that a plays against only η_1 workers is even lower. Considering only the marginal effect of effort on winning, then, a small prize will cause worker a to raise p_2 more than p_1. Increasing the performance levels, however, reduces $\mathrm{E}_\eta\{V\}$, so the tournament has a demotivating effect via the first term on the left side for either η_1 or η_2. Since $f(\bar{x} - p_1) > f(\bar{x} - p_2)$ this demotivating effect will be greater when $\eta = \eta_1$.

The tournament is ineffective in motivating η_1 workers partly because winning is an event that occurs only in the tail of the distribution, where changes in performance will not significantly affect the overall probability of winning. This argument is most compelling with a large number of workers and a concentrated prize structure (only one winner). The problem can be mitigated by reducing the concentration and size of the prizes. But in order to structure the tournament so that a bad η does not significantly disadvantage the worker, the prize would have to be substantially diluted (for example, prizes for the top 15 workers in a 20 person firm), reducing the overall incentive effect.

This leads to another useful way to understand low-powered incentives. An o-ring firm with heterogeneous performance needs to focus on motivating the bottom of the performance distribution. This is where a tournament with a concentrated prize structure is least desirable. Low-powered incentives of the type we study in this paper are, in effect, an inverted tournament in which virtually everyone is a winner. The only losers are those with very low subjective performance evaluation ($x < \bar{x}$). The logic is a mirror image of our analysis of a rank-order tournament (except that the firm fires a random rather than fixed number of workers). The prospect of losing their job has little effect on the η_2 workers; by drawing η_2 the deck is stacked against this outcome. In contrast, the η_1 workers realize that they can substantially affect the probability of avoiding this negative prize by supplying higher effort. Thus the low-powered scheme provides a relatively high level of motivation for η_1 workers with poor draws, while only minimally distorting the already acceptable performance levels of η_2 workers.

V. Conclusion

The arrangements employers typically reach with their workforce look quite different than incentive contracts derived by economic theorists. Many employees are hired with an implicit understanding that they will be offered a fixed level of compensation, along with continued employment, so long as their performance appears to exceed some minimal threshold. Thus Prendergast (1999) argues, "A critical avenue for future research should be to better understand the evaluation and compensation of those with non-contracted output."

These empirical observations motivate us to study low-powered incentives. In particular, we construct an environment in which the manager has a signal that serves as a subjective (private) summary statistic of the worker's performance. A subjective evaluation, based on human observation, has the advantage that it can extract information that is difficult to explicitly define or even articulate—politeness, enthusiasm, or cooperativeness—but which can be nonetheless be informally assessed by an observant manager. Obviously, firms cannot easily use such a signal as the basis of an incentive contract, but the signal can be used by the manager for making retention decisions.

Our theory provides a parsimonious characterization of compensation, performance, and turnover policy for a firm that uses low-powered incentives to solve agency problems. In the simplest version of our model, optimal low-powered incentives always induce the efficient performance level from employees. The retention rate does not depend on the dispersion of the firm's signal about worker performance, but compensation is linearly increasing in the variance. Motivation from our low-powered incentive stems from workers' desire to retain valuable jobs; the participation constraint is not typically binding. An explicit incentive, in the form of a rank-order tournament, may be particularly difficult to implement in an environment in which decisions are based solely on signals that are private information.

We find, though, that if the firm can implement such higher-powered incentives, they move workers closer to the participation constraint (with the residual agency problem solved by low-powered incentives).

Our most interesting results derive from a version of the model in which we assume that there is unobservable variation in the effort workers must supply to achieve a particular level of performance. In this instance, the efficiency of any incentive scheme hinges crucially on the extent to which there is complementarity in the performance level of the firm's workers. Extreme complementarity—Kremer's o-ring technology—implies that the value of the firm's output varies significantly with the poorest performance in the workforce. Low-powered incentives of the form we study fare especially well in this environment, because they operate so as to effectively motivate workers who would otherwise be most inclined to provide the lowest performance. Higher-powered incentives tend to squander resources on incentives that induce excess performance from workers already inclined to perform at high levels.

In most work environments, low-powered incentives have two further advantages (which we do not model in this paper) that doubtless tend to reinforce their use. First, if there is persistent heterogeneity in the quality of workers, the use of layoffs for particularly poor observed performance will tend to weed out poor workers. In this sense, the layoffs are doing double duty as an incentive device and a means of screening. Second, when low-powered incentives are used, workers who are penalized for poor performance will no longer be with the firm. Bewley (forthcoming) found that managers often defend their use of layoffs by noting that this "ships out" workers who would otherwise have the lowest morale. Understanding these two facets of a low-powered incentives policy would, minimally, require careful treatment of persistent heterogeneity and the attendant statistical learning issues.

Our work here persuades us that additional research on these lines might well

be fruitful. We believe, in particular, that research focusing on the interaction of incentives and the nature of the production technology holds promise for further theoretical development, and ultimately empirical testing.

Appendix

Proofs of Propositions 5 and 6

Proposition 5:

$$\theta g'(p_1^*) + (1-\theta)g'(p_2^*) = \theta\eta_1 + (1-\theta)\eta_2 = 1. \tag{11}$$

Proof: The firm's profit maximization is

$$\max_{\psi} \theta g(p_1) + (1-\theta)g(p_2) - w,$$

subject to equations (9). Note that (10) defines $\hat{p}_2 = p_2(\hat{p}_1, \bar{x})$ with

$$\frac{\partial p_2(p_1, \bar{x})}{\partial \hat{p}_1} = 1 - \frac{\partial p_2(p_1, \bar{x})}{\partial \bar{x}}, \tag{16}$$

which shortly proves convenient. Thus

$$\begin{aligned} w(\hat{p}_1, \bar{x}) = {} & \theta\eta_1\hat{p}_1 + (1-\theta)\eta_2 p_2(\hat{p}_1, \bar{x}) + (1-\beta)V^a \\ & + \frac{1-\beta[1-\theta F(\bar{x}-\hat{p}_1) - (1-\theta)F(\bar{x}-p_2(\hat{p}_1,\bar{x}))]}{\beta}\left(\frac{\eta_1}{f(\bar{x}-\hat{p}_1)}\right) \end{aligned} \tag{17}$$

defines the minimum w required to induce performance levels $\hat{p}_1$ and $p_2(\hat{p}_1, \bar{x})$ for a given $\bar{x}$ (as well as implicitly giving the best-response function $\hat{p}_1(\psi)$). An equivalent profit maximization is, therefore,

$$\max_{\bar{x}, \hat{p}_1} \ \theta g\big(\hat{p}_1\big) + (1-\theta)g\big(p_2(\hat{p}_1, \bar{x})\big) - w(\hat{p}_1, \bar{x}).$$

The first-order conditions are

$$\begin{aligned} 0 &= \theta g'\big(p_1^*\big) + (1-\theta)g'\big(p_2(p_1^*, \bar{x}^*)\big)\frac{\partial p_2(p_1^*, \bar{x}^*)}{\partial \hat{p}_1} - \frac{\partial w(p_1^*, \bar{x}^*)}{\partial \hat{p}_1} \\ 0 &= (1-\theta)g'\big(p_2(p_1^*, \bar{x}^*)\big)\frac{\partial p_2(p_1^*, \bar{x}^*)}{\partial \bar{x}} - \frac{\partial w(p_1^*, \bar{x}^*)}{\partial \bar{x}} \end{aligned}$$

Applying the envelope theorem, as in the proof of Proposition 2, and using (16), we find that

$$\frac{\partial w(p_1, \bar{x})}{\partial \hat{p}_1} + \frac{\partial w(p_1, \bar{x})}{\partial \bar{x}} = \theta\eta_1 + (1-\theta)\eta_2 = 1.$$

Proposition 5 follows from adding the two first-order conditions. ∎

Proposition 6: *Suppose that $\widetilde{g}(p)$ is closer to o-ring than $g(p)$. Then $\widetilde{p}_1 > p_1$.*

Proof: Suppose instead that $\widetilde{p}_1 \leq p_1$. This does not restore (11), that is,

$$\theta\widetilde{g}'(\widetilde{p}_1) + (1-\theta)\widetilde{g}'(p_2) > 1.$$

Therefore, if $\widetilde{p}_1 \leq p_1$, then $\widetilde{p}_2 > p_2$. Let ψ^* and $\widetilde{\psi}^*$ be the optimal employment policies for production functions g and $\widetilde{g}$, respectively. Let $\Pi(\psi)$ and $\widetilde{\Pi}(\psi)$ denote the profit functions in these two cases. Since the worker's best-response functions do not depend on g in any way, the firm can elicit performance levels $\widetilde{p}_1 = p_1(\widetilde{\psi}^*)$ and $\widetilde{p}_2 = p_2(\widetilde{\psi}^*)$ when the production function is g and $p_1 = p_1(\psi^*)$ and $p_2 = p_2(\psi^*)$ when the production function is $\widetilde{g}$. By assumption we have

$$\Pi(\psi^*) - \Pi(\widetilde{\psi}^*) > 0$$
$$\widetilde{\Pi}(\widetilde{\psi}^*) - \widetilde{\Pi}(\psi^*) > 0.$$

Adding these inequalities yields

$$\Pi(\psi^*) - \Pi(\widetilde{\psi}^*) + \widetilde{\Pi}(\widetilde{\psi}^*) - \widetilde{\Pi}(\psi^*) > 0.$$

With some rearranging this inequality becomes

$$\theta\Big(\big[g(p_1^*) - g(\widetilde{p}_1^*)\big] - \big[\widetilde{g}(p_1^*) - \widetilde{g}(\widetilde{p}_1^*)\big]\Big) + (1-\theta)\Big(\big[\widetilde{g}(\widetilde{p}_2^*) - \widetilde{g}(p_2^*)\big] - \big[g(\widetilde{p}_2^*) - g(p_2^*)\big]\Big) > 0$$

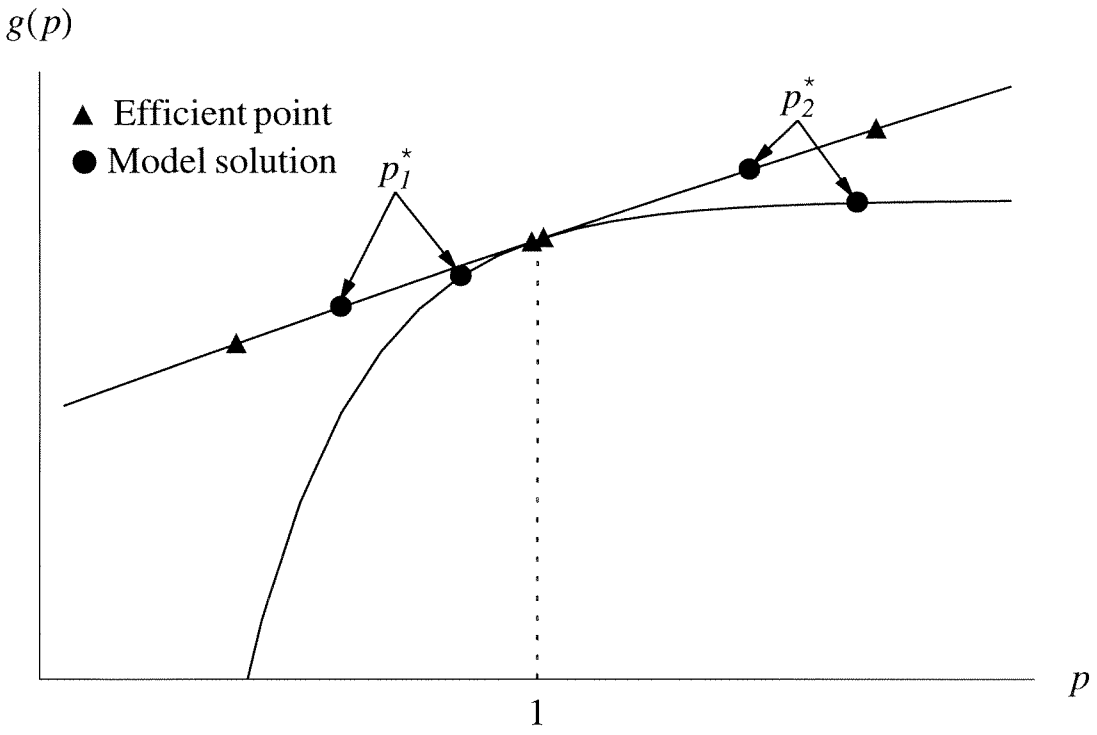
Given $\widetilde{p}_1 < p_1$ and $\widetilde{p}_2 > p_2$, each bracketed term is positive. Under our assumptions about the relationship between g and $\widetilde{g}$, however,

$$g(p_1^*) - g(\widetilde{p}_1^*) < \widetilde{g}(p_1^*) - \widetilde{g}(\widetilde{p}_1^*) \quad \text{and} \quad \widetilde{g}(\widetilde{p}_2^*) - \widetilde{g}(p_2^*) < g(\widetilde{p}_2^*) - g(p_2^*),$$

producing a contradiction. ∎

Figure 1

O-RING TECHNOLOGY AND UNOBSERVABLE HETEROGENEITY



References

Baker, George, Robert Gibbons, and Kevin J. Murphy. "Subjective Performance Measures and Optimal Incentive Contracts," *Quarterly Journal of Economics,* November 1994, 109(4), pp. 1125-56.

Bewley, Truman. *Why Not Listen to Business? A Study of Wage Rigidity,* forthcoming, Harvard University Press.

Gibbons, Robert. "Incentives in Organizations," *Journal of Economic Perspectives,* Fall 1998, 12(4), pp. 115-32

Holmstrom, Bengt. "Moral Hazard in Teams," *Bell Journal of Economics,* Autumn 1982, 13(2), pp. 324-40.

Holmstrom, Bengt and Paul Milgrom. "The Firm as an Incentive System," *American Economic Review,* September 1994, 84(4), pp. 972-91.

Kremer, Michael. "The O-Ring Theory of Economic Development," *Quarterly Journal of Economics,* August 1993, 108(3), pp. 551-75.

Kremer, Michael, and Eric Maskin. "Wage Inequality and Segregation by Skill," National Bureau of Economic Research Working Paper 5718, August 1996.

Lazear, Edward P. "Pay Equality and Industrial Politics," *Journal of Political Economy,* June 1989, 97(3), pp. 561-80.

Lazear, Edward P. and Sherwin Rosen. "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy,* October 1981, 89(5), pp. 841-64.

MacLeod, W. Bentley and Daniel Parent. "Job Characteristics and the Form of Compensation," 1997, mimeo, University of Southern California.

Malcomson, James M. "Work Incentives, Hierarchy, and Internal Labor Markets," *Journal of Political Economy,* June 1984, 92(3), pp. 486-507.

Mehta, Shailendra Raj. "The Law of One Price and a Theory of the Firm: A Ricardian Perspective on Interindustry Wages," *Rand Journal of Economics,* Spring 1998, 29(1), pp. 137-56.

Milgrom, Paul R. "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics,* Autumn 1981, 12(2), pp. 380-91.

Prendergast, Canice. "The Provision of Incentives in Firms," *Journal of Economic Literature,* March 1999, 37(1), pp. 7-63.

Prendergast, Canice and Robert H. Topel. "Favoritism in Organizations," *Journal of Political Economy,* October 1996, 104(5), pp. 958-78.

Shapiro, C. and J. Stiglitz. "Involuntary Unemployment as a Worker Discipline Device," *American Economic Review,* June 1984, 74(3) pp. 433-44.