The Principal’s Moral Hazard: Constraints on the Use of Incentives in Hierarchy

Gary J. Miller
Washington University

Andrew B. Whitford
University of Georgia

ABSTRACT

Pure incentive schemes rely on the agent’s self-interest, rather than more coercive control, to motivate subordinates. Yet most organizations, and in particular public agencies, rely very little on pure incentive contracts and instead use coercive mechanisms of monitoring and sanctioning that many theorists find objectionable. We use principal-agency theory to investigate the problem. Principal-agency theory has tacitly assumed throughout that it is in the principal’s interests to find a set of incentives that induce efficient levels of effort from the agent. We show that this is not necessarily the case. We identify a problem we denote as “the principal’s moral hazard constraint” in which bonuses large enough to produce the efficient incentive effect are prohibitively expensive for the principal. Potential solutions to this problem—involving penalization or joint ownership—are unavailable in the public sphere. This means that for a large class of control problems in agencies, the principal’s self-interest will result in the inefficient use of monitoring and oversight rather than outcome-contingent incentives. Although monitoring is often thought of as resulting from the agent’s moral hazard, it can just as reasonably be seen as resulting from the principal’s moral hazard.

INTRODUCTION

Since Adam Smith, the attraction of the market has been the Invisible Hand: the promise that social well-being may be advanced simply by allowing individuals to follow their self-interest. This benign result is made possible through the price mechanism, which is driven by self-interest to efficient allocations of resources (assuming conditions regarding competition, externalities, and information). With no need for coercion, the market requires no difficult trade-off between social efficiency and the moral benefits of liberty. Because the price mechanism does not operate in a hierarchy (Coase 1937), economists traditionally had little to say about the more coercive hierarchical institutions of society, except to regret the necessity of their existence.

More recently, principal-agency theory has led to an explosion of interest in the role of incentives that goes well beyond the price mechanism to look at incentives inside firms,
government agencies, and contracting in general (Harris and Raviv 1979; Eisenhardt 1989). The original dichotomy between the price mechanism and hierarchy has been abandoned. Works on principal-agency theory (Baron and Besanko 1984; Laffont and Martimort 2001) are rich with the optimism that efficiency gains may be realized within the firm and public bureaucracies by invoking self-interest under the right incentive contract.

Essentially, the central result in principal-agency theory is that even the confounding factor of asymmetric information can be overcome by the right incentives (Harris and Raviv 1979; Holmstrom 1979). The premise is that, even in hierarchical institutions, much of the work of controlling subordinate behavior can be left to the subordinate’s self-interest, guided by the correct incentives; the expense and moral ambiguity of monitoring, rule making, and coercion can be largely avoided.

This shift in thinking is pervasive: as a premier treatise on principal-agency theory notes, “Today, for many economists, economics is to a large extent a matter of incentives . . .” (Laffont and Martimort 2001, 1). Over time, the search for incentives has come to permeate the study of firms and government alike. In turn, the study of principal agency is a dynamic area of study in economics, with numerous studies attempting to test the predictions of agency theory (e.g., Prendergast 1999), often experimentally (Conlon and Parks 1990; McLean Parks and Conlon 1995; Fehr, Gächter, and Kirchsteiger 1997; Fehr and Falk 1999; Miller and Whitford 2000; Bottom et al. 2006). Agency theorists have examined repeated play among sets of individuals (Meyer and Vickers 1997), multiple agents (Macho-Stadler and Pérez-Castrillo 1993, 2001; Gupta and Romano 1998), and multiple levels of hierarchy (Masking, Qian, and Xu 2000).

Over the past three decades, theory has de-emphasized the power of monitoring and increased its emphasis on monetary incentives. Alchian and Demsetz (1972) argued that the principal’s role is to monitor and (potentially) chastise subordinates. Holmstrom (1982, 325–6) took the radically opposed position that, “the principal’s role is not essentially one of monitoring . . . . The primary role of the principal is to administer incentives schemes that police agents in a credible way rather than to monitor agents as in Alchian and Demsetz’ story.” The hope is that in the end, if the principal gets the incentives right, the organization will become a machine that runs by itself, fueled by an adequate supply of self-interested behavior. Managers will not have to invest heavily in thick rule books, close monitoring, organizational structures, or process control to make sure that subordinate behavior is in line with managerial expectations.

This often amounts to a belief that virtually every principal-agency relationship can be resolved by incentives, making an investment in hierarchical coercion unnecessary. Firms and governments alike have increased their reliance on markets by downsizing their own staff in favor of obtaining goods and services through external market contracts. For the remaining personnel in downsized firms and government agencies, the right incentives import the Invisible Hand from the marketplace to the firm; each organization member, consulting her incentives, finds it in her interest to do just what the organization needs for efficiency. This diminishes the need for investment in supervision and the use of sanctions—or so the argument runs.

In our view, the question is not “why use incentives?” but, given the incentives revolution, “why are incentives not used more often than they are?” A number of recent studies document the puzzling lack of incentives even in corporate hierarchies (e.g., Baker, Jensen, and Murphy 1988). They find that even in those cases where the agent’s performance can be directly linked to compensation, corporations often still choose to use
contingency-free compensation schemes. Furthermore, incentives are very seldom used in most governmental agencies, at any level. Even in the case of the Senior Executive Service (SES) (which was ostensibly formed to better link pay and performance) few contingencies are evident in actual compensation rates.

One potential explanation for this is risk aversion. Risk aversion, combined with information asymmetry, makes it impossible to replace monitoring of agent behavior with an equally efficient system of incentives based on easily observed outcomes (Harris and Raviv 1979; Holmstrom 1979). In this situation, the outcome-based incentives needed to guarantee efficient actions from the agent necessarily undermine the efficiency of risk sharing. Either risk is efficiently borne by the principal (which leaves the agent with an incentive to shirk) or the risk is inefficiently shifted to agents in order to create incentives that overcome moral hazard (Harris and Raviv 1979).

The purpose of this essay is to point out important logical limitations on incentives even when agents are risk acceptant. We argue that there are systematic reasons why self-interested principals will not choose to use efficient incentives, out of a concern for the principals’ own profits. This may go far to explain why firms—and especially, many public agencies—fail to impose incentives-based schemes. Specifically, there is a reasonable set of cases in which the bonus-based incentives necessary to induce efficient action by the agent are more costly to the principal than they are worth to her. That is, the use of incentives may be efficient but not profit maximizing. In effect, the stars must be aligned just exactly right in order for the use of incentives to be both efficient and profit maximizing. And in public agencies, we expect that often the stars are not so aligned.

Our purpose here is to document two simple constraints on the use of incentives. Both are framed in terms of the relationship between the agent’s efficacy—the agent’s ability to affect the desired outcome—and the agent’s outcome-based bonus. The first is a social efficiency constraint: when the agent’s efficacy is sufficiently low, the bonus necessary to induce socially desirable behavior is inefficiently large. In such a situation, the organization will have to hope that a monitoring and sanctioning system can be implemented because an outcome-based incentive system cannot.

Second, and more troubling, we identify a moral hazard constraint for principals—that even when an outcome-based incentive system is socially efficient, the principal may prefer not to implement it. The reason for this is a form of moral hazard: the principal can expect to earn a higher residual profit with an inefficient contract than if the agent is induced by a bonus to take the socially efficient outcome. The bonus necessary to induce the efficient action is simply greater than the possible return to the principal. The principal’s self-interest tempts the principal away from the efficient use of incentives.

This essay proceeds as follows: In the next section, we offer a simple model of the principal-agency relationship to illuminate the efficiency constraint on the use of incentives. If the agent’s efficacy is too low (in comparison to the marginal cost of effort required to achieve ex ante efficiency), it may not be possible to generate efficient outcomes by invoking the right contingent incentive scheme. In the third section, we extend this model to describe the principal’s binding rationality constraint. In the fourth section, we discuss two theoretical solutions or ways out of the principal’s trap: either to penalize the agent for poor performance or to form an ownership agreement with the agent. These solutions face accompanying constraints: the use of penalties faces legal restrictions in the public workforce and ownership-based schemes are impossible due to the ownership structure of government. This means that for a large class of bureaucratic control problems,
and hierarchical control problems more generally, institutional designers will not be able to rely on incentives to align principal and agent preferences—the Invisible Hand will not be able to overcome an inherent conflict of interest in the hierarchy. Last, we offer a discussion of this theory and its implications for a theory of democracy.

INCENTIVES AND THE CRITICAL LEVEL OF THE AGENT’S EFFICACY

The paradigmatic problem addressed by principal-agency theory is one in which an agent and a random variable interact to produce an outcome of value to the principal. For example, a tenant farmer’s crop is determined jointly by his own effort and the weather. In general, we assume the probability of a good crop increases with the farmer’s efforts. However, this fact does not allow the farm owner to deduce anything about the farmer’s efforts from the final outcome. If there is a bad crop one year, the farmer may blame the weather, even though he himself shirked. In a good year, he may take the credit, although the weather played a large part. This is the problem of information asymmetry that is basic to principal-agency theory.

The farm owner could make an investment in monitoring, so that he can pay the tenant farmer only if he works hard. However, this is usually a costly process. The problem, conceived of as the principal’s problem, is to design a contract that will induce the tenant farmer to work even without monitoring and sanctions. For example, sharecropping is a form of contract in which the tenant farmer comes to share, with the owner, a strong self-interest in a successful crop. The owner can then presume a high effort on the part of the farmer without ever having to verify it.

The same situation applies elsewhere. In a corporation, the principal is often thought of as the owner of a firm, and the agent is an employee with specialized expertise. In a public agency setting, we may think of the principal as a political figure (e.g., a mayor) and the agent as a specialized expert.

In the simplest model, the principal has an asset that is worth one of two possible values: \( S > F \geq 0 \). In the private setting, this could be thought of as ownership rights in a computer chess game, which will be worth either more \((S)\) or less \((F)\) depending on the quality of the programming supplied by the agent (programmer) (Dixit and Nalebuff 1993). In a public setting, a politician may run for reelection on the promise of a successful public safety campaign; \( S \) would then be the discounted stream of benefits from the likely reelection based on the success of the public safety campaign, whereas \( F \) would be the (smaller) discounted stream of benefits from a failure of the campaign. The successful outcome is more likely if the police chief (agent) supplies a high effort rather than a low effort. In the context of our model, the politician (principal) perceives that her own electoral benefits are in some way connected to the public goods program supplied by the bureaucrat (agent).

The value of the asset depends on both a random variable and agent effort. In this simple model, the agent may choose from three actions: HI, LO, and EXIT. If she chooses not to exit, the agent can either apply herself diligently to the programming tasks or not. If the agent chooses EXIT, then the asset is worth zero with certainty. A choice of HI effort produces a greater probability of success than a LO effort. The probability of \( S \) given HI is \( p \); the probability of success given LO is \( q \) (with \( p > q \)). Both the private firm owner and the politician/principal would like, other things being equal, to invoke a high effort from the agent because a high effort increases the probability of \( S \) instead of \( F \). The programmer or police chief can apply herself diligently or just go through the motions.
If the agent’s actions were directly and cheaply visible to the principal, there would be no problem: the principal could just write a contract in which the agent is paid only for a high effort. But, in those interesting and important situations present in most hierarchies, this is not possible. The difference in effort is not visible to the principal; that is, the agent’s actions are important for the principal’s ultimate payoff but are hidden from the principal’s direct knowledge. The question is whether the principal can come up with a set of incentives that will induce a HI effort, despite the agent’s “hidden” actions. Can incentives efficiently replace information?

Our first observation is that the answer to this question depends in a large part on the agent’s efficacy. If \( p \) is substantially greater than \( q \), then the agent is highly effective—the agent can make a substantial difference in the probability of success by supplying HI rather than LO effort. For example, the programmer could have a very large impact on whether the computer game is a success, or the police chief could have a large impact on whether the crime reduction program is a success. On the other hand, if \( p - q \) is close to zero, then the agent is ineffective. We use the term agent efficacy to refer to the quantity \( p / q \). We will shortly demonstrate that a lower level of agent efficacy decreases the likelihood that incentives can profitably be used to motivate the agent.

The agent experiences a personal effort cost, \( C \), of these actions, with \( C_{HI} > C_{LO} \geq C_{EXIT} = 0 \). As with agent efficacy, the difference between \( C_{HI} \) and \( C_{LO} \) is crucial to the analysis. If the cost of a HI effort is close to the cost of a LO effort, then it is relatively easy to induce the agent to take the HI effort, thereby increasing the chance of success. If the two cost terms are quite different, then the incentives necessary to induce a HI effort will of course be greater. We call \( C_{HI} - C_{LO} \) the marginal cost of HI effort.

As is standard in principal-agency theory, we assume that the agent’s actions are prohibitively costly to detect, so that the principal and agent can contract on the outcome but not on the agent’s actions. The contract can include a flat wage, \( W \), and an outcome-contingent bonus, \( B \), paid only in the event of a success, \( S \). The agent wants to maximize compensation net of the effort cost. The principal seeks to maximize the value of the asset net of the payment to the agent. This model is roughly the same as described in Laffont and Martimort (2001, chap. 4); it is the basis for conclusions in Dixit and Nalebuff (1993), Miller and Whitford (2000), and Bottom et al. (2006).

Clearly, the use of an outcome-contingent bonus is less likely in a public setting, but at this stage, we are considering all possibilities, so we explore the public and private principal-agent relationships in a parallel way. At this point, all we need is that there are some private and public principals (read, private asset owner and public good entrepreneur) who have some stake in the hidden actions of expert subordinates.

We want to be clear that our approach hinges on an agent’s behavior not being observable. The principal has no “ability to control those actions that are no longer observable, either by the principal who offers the contract or by the court of law that enforces it. Those actions cannot be contracted upon because no one can verify their value” (Laffont and Martimort 2001, 145). Of course, not all public settings are marked by perfect unobservability. As J. Q. Wilson (1989, 158) argued, for some agencies the activities of operators can be observed and for some agencies the results of those operators’ activities can be observed; Wilson’s four agency types—production, procedural, craft, and coping—correspond to different combinations of observable actions and outcomes. In the theory we offer here, the assumptions of classic principal-agency theory correspond most closely to Wilson’s notion of craft organizations, outcomes are observable and actions are not. However, we emphasize
that even in production and procedural organizations outputs are not perfectly observable. In these cases, some reliance on outcome-contingent contracts may also be worthwhile.

To emphasize that the constraints discussed in this essay are different from the normal problems associated with risk aversion, we assume that both the principal and agent are risk neutral. The standard analysis is that with risk-neutral agents, there is no efficiency loss from information asymmetries. As Harris and Raviv (1979, 223) conclude, incentives are perfectly effective with risk-neutral agents: “There are no gains to be derived from monitoring the agent’s action when the agent is risk neutral.” We will show that, for practical purposes, an organization staffed with risk-neutral employees may nevertheless end up employing monitoring, rule making, directives, and coercive sanctions—either through an efficiency constraint or through the moral hazard of the principal.

The first thing to notice is that though a flat wage may help keep an agent from exiting the relationship, it does nothing to motivate a HI effort rather than LO. The expected value of HI and LO effort both go up by the amount of the fixed wage. In order to differentially motivate a HI effort, the expected value of the HI effort must go up more than the expected value of LO effort. This is accomplished by the promise of a bonus that is contingent on outcomes: $S$ versus $F$. As long as a HI effort makes $S$ more likely, then the bonus can help to motivate a HI effort.

The effective bonus $B^*$ will increase the agent’s expected net benefit with a HI effort so that it is greater than that with a LO effort:

$$pB^* + W - C_{HI} \geq qB^* + W - C_{LO}.$$  

This can be written as

$$B^* \geq \frac{C_{HI} - C_{LO}}{p - q}.$$  

The right-hand term is just the ratio of the marginal cost of HI effort divided by the agent’s efficacy. A large $p - q$ implies that the agent’s choice of HI or LO has a greater impact on the principal’s well-being and on his own chances of a bonus. The lower the agent’s efficacy, the higher the critical bonus must be (holding other parameters constant). This is necessary to offset the lower gains attributable to the outcome that probabilistically results from HI effort rather than LO.

For example, consider the following parameters: The asset may be worth $30,000 (in our model, $S$) or $10,000 (in our model, $F$). The cost of a LO effort is half that of a high effort: $C_{LO} = $4,000 and $C_{HI} = $8,000. Let $p = 0.8$ and $q = 0.4$. In this case, $B^*$ must be

---

1. In canonical agency theory, risk aversion on the part of the agent is more troubling for designers of optimal incentives because risk aversion induces an “insurance/incentives” trade-off for which the design must account (Laffont and Martimort 2001). This trade-off has been empirically investigated through the use of experiments (e.g., Fehr and Falk 1999; Miller and Whitford 2000). We emphasize that our model involves linear utility functions for both the principal and the agent.

2. In addition, the bonus should be high enough to prevent the agent from exiting:

$$B^* \geq \frac{C_{HI}}{p}.$$  

This will be the case whenever $C_{LO} \geq C_{HI}$, which would be the case (for example) whenever shirking results in a very low probability of success.

3. One interpretation of these parameters is that the $S$ outcome is essentially a probabilistic “tripling” of the agent’s investment in a high effort.
at least: \( B^* \geq (8,000 - 4,000)/(0.8 - 0.4) = 10,000 \), which is a significant, but not impossible, bonus. With any bonus equal to \( B^* \) or greater, the risk-neutral agent would supply a HI effort simply as a result of her own self-interest, with no other inducement or hierarchical monitoring by the principal; because it induces high effort, the bonus would also make the principal better off. Thus, with these parameters, the standard conclusion from principal-agency theory applies: the use of an outcome-contingent bonus induces an efficient level of effort and is Pareto optimal. Investments in monitoring and oversight are costly, inefficient, and unnecessary. Incentives substitute perfectly for the principal’s missing information about the agent’s actions.

**EFFICIENCY CONSTRAINTS ON THE AGENT’S EFFICACY**

As agent efficacy decreases to a certain point, it is no longer efficient to try to induce HI. This point occurs when the extra social surplus generated by the extra effort is larger than the marginal effort cost of HI. The basic measure of the benefits of the principal-agent interaction is the expected revenue less the agent’s effort costs. The expected social surplus depends on whether the agent supplies a HI or LO effort. In the former case, the expected social surplus is equal to \( p(S) + (1 - p)F - C_{HI} \). If the agent supplies a LO effort, the expected social surplus is equal to \( q(S) + (1 - q)F - C_{LO} \).

The expected social surplus resulting from a HI effort is greater than that from a LO effort if and only if

\[
(p - q)(S - F) \geq C_{HI} - C_{LO}.
\]

This efficiency constraint can be rewritten as a constraint on agent efficacy \( (p - q) \):

\[
p - q \geq \frac{C_{HI} - C_{LO}}{S - F}.
\]

This means that, for given asset values and effort costs, there is a critical level of efficacy \( (p - q) \) below which incentives are inefficient. That is, the use of an incentive bonus does nothing but generate an inefficiently high level of effort. With \( S = 30,000 \), \( F = 10,000 \), and the cost parameters from the original example in the previous paragraph, then \( p - q \) must be at least equal to 0.2 in order to justify the use of the minimal incentive bonus \( B^* \).

\[
p - q \geq \frac{8,000 - 4,000}{30,000 - 10,000} = 0.2.
\]

Whenever \( p - q < 0.2 \), then the extra $4,000 in effort costs associated with a HI effort cannot result in an extra $4,000 of expected value.

This efficiency constraint on incentives is shown in figure 1, which assumes that the parameters are as in the previous example, with the exception of \( q \), which is presumed to vary from 0.8 to 0.0 (from left to right). Since \( p = 0.8 \), this variation in \( q \) produces an efficacy variable \( (p - q) \) from 0.0 to 0.8 (reading from left to right). The expected social surplus given a HI effort is constant at $18,000 (since it is not a function of \( q \)). The expected social surplus given a LO effort starts out $4,000 greater than expected social surplus/HI because effort costs are $4,000 less. The expected social surplus decreases as \( q \) increases. The two lines cross at agent efficacy equal to 0.2. Only when the agent’s marginal impact on the probability of a success is greater than 0.2 is it efficient to pay the extra $4,000 to induce a high effort.
This possibility clearly has real-world significance. That is, most jobs are ones in which it is simply inefficient to provide outcome-contingent incentives for success. If the individual police officer’s effect on the crime rate is largely determined by economic, social, and political factors beyond the control of the individual police officer, then $p - q$ would be low and the bonus necessary to motivate high effort by a police officer would be prohibitively expensive. If a police officer has only a 1% probability of making a crime reduction program a success, and the marginal effort cost is even quite modest (at $10,000), then $B^*$ would be $100,000. The alternative, once again, is to hire a supervisor (e.g., a police sergeant) who tries to detect whether the police officer is in fact providing full effort and sanction the officer who is found shirking. A police sergeant who earns less than $100,000 and can supervise 8–10 police officers is a necessary alternative to the expense of bonuses.

For the rest of the essay, however, we look at the special category of jobs for which incentive bonuses are efficient because equation (4) is satisfied. We ask whether the principal’s self-interest will lead her to impose outcome-contingent incentives, even when they are known to be efficient. We will find that the principal can often generate higher expected profits for herself by failing to impose efficient incentives.

**THE PRINCIPAL’S CALCULUS**

Given these same asset values ($S = $30,000 and $F = $10,000) and cost parameters ($8,000 for a HI and $4,000 for a LO effort), consider a situation in which $p = 0.8$ and $q = 0.55$. In this case and with these parameters, it is efficient to use a bonus to induce high effort: $B^* \geq (8,000 - 4,000)/(0.8 - 0.55) = $16,000. With a LO effort, the expected value of the asset is $EV_{LO} = qS + (1 - q)F = $21,000; subtracting $C_{LO}$ yields an expected social surplus of $17,000. With a HI effort, the expected value of the asset is $EV_{HI} = pS + (1 - p)F = $26,000; subtracting $C_{HI}$ yields a greater expected social surplus of $18,000. This means that the socially efficient outcome is to induce a HI effort.
The principal can induce a HI effort by offering a bonus, \( B^* \), of $16,000; any bonus less than that value would not be sufficient to compensate the agent for the extra cost of HI effort. This bonus leaves the principal with an expected profit of
\[
\pi = p(S - B^*) + (1 - p)F = \$13,200.
\]

Can the principal find a way to produce an expected profit larger than $13,200? Suppose the principal decides not to use a bonus, knowing full well that the agent will simply supply a LO effort. The principal can induce a LO effort by a flat wage sufficient to cover the cost of low effort: $4,000. Just as we use \( B^* \) to denote the minimum bonus necessary to induce a HI effort, we use \( W^* \) as the minimum wage necessary to induce a low effort (to prevent the agent from quitting altogether.) The use of \( W^* \) offers a significant savings in agent compensation. In fact, the principal’s expected profit is
\[
\pi = q(S - C_{LO}) + (1 - q)(F - C_{LO}) = \$17,000.
\]
The principal’s use of flat wage induces an inefficient LO effort but greater profits.

The use of the flat wage increases the principal’s profit, even though the expected social surplus decreases from $18,000 to $17,000. The principal faces a form of moral hazard because she expects more profit by inducing a low effort than a high effort even though a high effort is clearly efficient. The principal has every incentive to sabotage the efficient outcome since she wishes to minimize the payment to the agent.

We believe that this is a significant general point. Much of the principal-agency theory has taken the principal’s point of view and has treated the maximization of the principal’s profits as being the same thing as efficiency. However, efficiency (maximizing the size of the pie to be divided between principal and agent) is not the same thing as maximizing the principal’s share of the pie. As long as the choice is restricted to a positive flat wage or a positive bonus, there is no incentive scheme that can in general reconcile the requirements of organizational efficiency and the principal’s self-interest. There is no bonus less than $16,000 that will induce a HI effort, but a bonus greater than $16,000 diminishes the principal’s expected profits. The principal cannot keep for herself enough of the extra expected revenue to cause her to motivate a high effort because in order to gain the agent’s effort the revenue must flow toward the agent’s bonus. In general, then, the principal may prefer not to take the action that generates the extra social surplus gained by inducing the efficient level of effort. This drives a sharp wedge between social efficiency and principal rationality. A self-interested solution to the “principal’s problem” will often entail ignoring efficiency.

PROFITABILITY CONSTRAINT ON THE AGENT’S EFFICACY

We can show that the profitability constraint, like the efficiency constraint, can be written as a constraint on the agent’s efficacy. The agent’s efficacy must be higher than a certain level to make it worth the principal’s while to pay the bonus necessary to induce HI effort. Further, the profitability constraint is higher than the efficiency constraint, which implies that there is a midrange of efficacy levels in which the principal’s self-interest is the obstacle to the use of efficient incentives.

The principal’s expected cost of inducing high effort with a bonus \( B^* \) is \( pB^* \), the expected cost of inducing a flat wage is \( C_{LO} \), and the expected marginal revenue is
The principal’s individual rationality constraint for the use of the bonus $B^*$ is

$$pB^* - C_{LO} < (p - q)(S - F)$$

or

$$B^* < \frac{(p - q)(S - F) + C_{LO}}{p}. \quad (6)$$

The upper bound for the region in which the principal will select a Pareto inefficient contract can be found by substituting from equation (2) into equation (6), stated as an equality, and then solving for $p - q$ with the quadratic equation. The minimum level of $p - q$ for a given $p$ at which feasible contracts can be written is

$$p - q = \frac{C_{LO} + \sqrt{4p(S - F)(C_{HI} - C_{LO}) + C_{LO}^2}}{-2(S - F)}. \quad (7)$$

For the parameters offered here, $(p - q) = 0.3125$. Combined with the result in (4a), this gives a range in which the principal experiences moral hazard: the incentive solution to the principal’s problem is efficient but not chosen by the principal when $0.2 < (p - q) < 0.31$. Empirically, this means that we should see efficient incentives chosen by principals only when the agent’s efficacy is especially high.

This is illustrated in figure 2, which includes the expected social surplus graphs from figure 1 and adds the expected profit graphs resulting from contracting $B^*$ and $W^*$. All parameters except $q$ are as in cases 1–3 in table 1 (see subsequently). The variable $q$ (the probability of success when shirking) ranges from 0.8 to 0.0 (it must be less than 0.8 because if it was greater the agent could shirk and have a more positive impact on the outcome than by providing a HI effort). As in figure 1, a HI level of effort is efficient for efficacy greater than 0.2.
We notice that the expected profit given $W^*$ is the same as the expected social surplus given $W^*$ (it is the same linearly decreasing line). The reason for this is that, as long as the principal is satisfied with a LO effort, the contract based on $W^*$ minimally reimburses the agent for the LO effort cost and results in virtually the entire social surplus staying in the hands of the principal. As a result, the expected profit decreases with efficacy just as expected social surplus does, given $W^*$. At the right margin of the figure, expected profits with $W^*$ are $6,000 because the owner has a near certainty of an asset worth $F = 10,000$, at a labor cost of $4,000$.

As $q$ decreases, $p - q$ (agent efficacy) increases, and over this range, $B^*$ decreases from an infinite number at $(p - q) = 0$ to $20,000$ at $(p - q) = 0.2$ to $10,000$ at $(p - q) = 0.4$ to $4,000$ at $(p - q) = 0.8$. The decrease in the size of the bonus with efficacy means that the owner’s expected profits increase with efficacy.

However, we see that expected profits given $W^*$ dominate expected profits given $B^*$ until $(p - q) = 0.31$. This is because when agent efficacy is moderate, the size of the bonus necessary to induce HI effort is large. The principal’s profit given $B^*$ starts off in the negative region, rises rapidly with agent efficacy (as $B^*$ decreases), and then flattens out. It does not cross the graph of expected profit with $W^*$ until efficacy is approximately equal to 0.31. In the range of efficacy from 0.2 to 0.31, efficiency is maximized with the use of $B^*$, but the principal herself achieves higher profits using a low flat wage, $W^*$.

The rows of table 1 show the efficiency constraints and profitability constraints on the agent’s efficacy, given a range of contracts. Cases 1–3 use the parameters from the discussion above and are noted on the “agent efficacy” axis of figure 2. In case 1, which we presented above, the agent’s minimum necessary efficacy for responding to an incentives-based contract is 0.2; the agent’s actual efficacy is 0.25. However, the principal will not find it profitable to use incentives unless the agent’s efficacy is at least 0.31; the principal shirks. In case 2, the agent’s actual efficacy is 0.1; the agent is unwilling to respond to an incentives-based contract with high effort, and the principal is unwilling to offer such a contract. In case 3, the agent’s actual efficacy is 0.6: it is rational for the principal to offer an incentives-based contract and the agent is willing to respond to such a contract.

<table>
<thead>
<tr>
<th>Case</th>
<th>$S$ (in $)</th>
<th>$F$ (in $)</th>
<th>$p$</th>
<th>$q$</th>
<th>$C_{HI}$ (in $)</th>
<th>$C_{LO}$ (in $)</th>
<th>Agent efficacy $(p - q)$</th>
<th>Efficiency constraint on agent efficacy</th>
<th>Profitability constraint on agent efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.55</td>
<td>8,000</td>
<td>4,000</td>
<td>0.25</td>
<td>0.20</td>
<td>0.31</td>
</tr>
<tr>
<td>2</td>
<td>30,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.70</td>
<td>8,000</td>
<td>4,000</td>
<td>0.10</td>
<td>0.20</td>
<td>0.31</td>
</tr>
<tr>
<td>3</td>
<td>30,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.20</td>
<td>8,000</td>
<td>4,000</td>
<td>0.60</td>
<td>0.20</td>
<td>0.31</td>
</tr>
<tr>
<td>4</td>
<td>30,000</td>
<td>10,000</td>
<td>0.90</td>
<td>0.55</td>
<td>8,000</td>
<td>4,000</td>
<td>0.35</td>
<td>0.20</td>
<td>0.34</td>
</tr>
<tr>
<td>5</td>
<td>30,000</td>
<td>10,000</td>
<td>0.60</td>
<td>0.55</td>
<td>8,000</td>
<td>4,000</td>
<td>0.05</td>
<td>0.20</td>
<td>0.26</td>
</tr>
<tr>
<td>6</td>
<td>40,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.55</td>
<td>8,000</td>
<td>4,000</td>
<td>0.25</td>
<td>0.13</td>
<td>0.27</td>
</tr>
<tr>
<td>7</td>
<td>20,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.55</td>
<td>8,000</td>
<td>4,000</td>
<td>0.25</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td>8</td>
<td>30,000</td>
<td>5,000</td>
<td>0.80</td>
<td>0.55</td>
<td>8,000</td>
<td>4,000</td>
<td>0.25</td>
<td>0.16</td>
<td>0.29</td>
</tr>
<tr>
<td>9</td>
<td>30,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.55</td>
<td>10,000</td>
<td>4,000</td>
<td>0.25</td>
<td>0.30</td>
<td>0.40</td>
</tr>
<tr>
<td>10</td>
<td>30,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.55</td>
<td>6,000</td>
<td>4,000</td>
<td>0.25</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>11</td>
<td>30,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.55</td>
<td>8,000</td>
<td>6,000</td>
<td>0.25</td>
<td>0.10</td>
<td>0.17</td>
</tr>
<tr>
<td>12</td>
<td>30,000</td>
<td>10,000</td>
<td>0.80</td>
<td>0.55</td>
<td>8,000</td>
<td>2,000</td>
<td>0.25</td>
<td>0.30</td>
<td>0.44</td>
</tr>
</tbody>
</table>
with high effort. Cases 1–3 involve manipulation of only the parameter \(q\); this means that variation in the agent’s impact of low effort shifts the agent across a stable range of feasible and infeasible contracts (from the principal’s perspective).

Cases 4 and 5 show that changing the agent’s impact of high effort increases the principal’s moral hazard range: that is, as \(p\) gets larger, the efficacy necessary for profitable use of incentives diverges from that necessary for the efficient use of incentives. Cases 6 and 7 show that changing \(S\), the value of the asset in the case of a success, changes both the efficiency and profitability constraints on efficacy. Case 8 shows similar changes in the case of \(F\). Cases 9 and 10 show the changes in the principal’s moral hazard given variation in \(C_{HI}\); cases 11 and 12 show the changes given variation in \(C_{LO}\).

The last two columns of table 1 indicate the range of the principal’s moral hazard. The level of agent efficacy that is necessary for profitable use of contingent incentives is greater than the level of agent efficacy necessary for the efficient use of contingent incentives. When \(p - q\) lies in this range, the social surplus would be maximized by the use of a contingent bonus—but the bonus necessary to induce a high effort by the agent is too large to be profitable for the principal. The principal would prefer to maximize profits by inefficiently satisficing with a flat wage and the resulting low effort.

As cases 1–3 indicate, the moral hazard range is fixed by the parameters other than \(q\) (the impact of low effort). This means that \(q\) is a free variable in the analysis—by changing \(q\) only, we can alter agent efficacy, without altering the efficiency and profitability constraints. As cases 5–7 indicate, changing other parameters can expand or contract the range of the principal’s moral hazard.

**INEFFICIENT MONITORING**

The inability to claim the efficiency gains generated by incentives may result in either inefficiently low effort or an inefficient reliance on monitoring to attain high effort. To continue the example given above, consider if the services of a monitor were available for $750. With this monitor, the principal could pay the agent $8,000 to cover the cost of HI effort, pay the monitor $750, and get a higher expected profit ($17,250) than by either using the efficient bonus or by acceding to LO effort. But this, too, is a form of inefficiency because the payment to the monitor is a deadweight loss.

Thus, whenever agent efficacy \((p - q)\) is sufficiently low, then we should expect self-interested principals either to adopt a monitoring plan or to impose flat wage levels consistent with shirking. In fact, empirical research on firms, specifically in the case of sales, seems to support many of these results. For example, Eisenhardt (1989) found that routine sales jobs, such as operating a cash register, were paid with fixed salaries, whereas sales positions that involved establishing a close relationship with the customer were paid with a commission. This is consistent with principal-agency theory in that employers bore the risk for those positions in which employees could be cheaply and easily monitored, but risk was shifted to those employees where monitoring was more expensive. Presumably, agents with the most risk aversion were more likely to stick with lower paid but safer fixed wages, operating the cash register, whereas agents who had the least risk aversion were those who accepted the commission pay for the less easily monitored personal sales positions. Indeed, past studies of sales compensation plans have revealed that most plans include a very substantial fixed pay component that is not consistent with the agency theory (Eisenhardt 1989; Coughlan and Narasimhan 1992). Although
Aggarwal and Samwick (1999) argue that executive pay plans are moving toward more variable pay, and thus toward greater consistency with agency theory than traditional practices, this in fact may be a time-bound result, given recent changes in the economy and the changing use of stock options for incentivization. Of course, such field studies are also inevitably limited by a lack of control and randomization (Tosi et al. 2000). Specifically, an agent’s type and level of labor skills may be related to their degree of risk aversion. Our inferences from these studies are limited because of their inability to control for ex ante risk aversion.

In both public and private agencies, the situations in which agent efficacy is sufficiently high that incentives may be effectively implemented are rare. Even in private firms, it is rare that the link between individual effort and outcome is tight enough to justify the use of piece rates. The linkage may be even more tenuous in public bureaucracies, where Lipsky (1983) notes a deadly combination: ambitious and vague goals, together with limited resources. These, together with biased performance indicators lead to alienation from clients and goal displacement. Furthermore, the measurable outcomes that are a legitimate goal of public policy are normally impacted by a variety of essentially random variables beyond the control of the bureaucrat. The contractible outputs of a police patrol—crime rate on the cop’s beat, for example—are largely dictated by variables such as employment rate, parental presence, the actions of drug pushers, to name a few. Even if the public found it acceptable to give the individual patrol officer a bonus when the crime rate is significantly reduced, the probability that the individual officer could significantly increase the probability of earning the reward is likely to be small indeed. The size of the bonus that would be necessary to induce officers to significantly change patrolling behavior (by focusing on high-risk blocks or confronting violent criminals) would be extremely large. The public (the police officer’s ultimate principal) would balk at six-figure bonuses for officers even if those bonuses were efficiently small compared to the public benefits of reduction of crime.

Of course, students of public administration have long debated the efficacy of outcome-based incentives in organizations, in part due to the historical role of Barnard’s (1938) writings on authority in complex organizations and Simon’s (1945) reconstruction of incentives to include a broader set of considerations. In fact, little empirical evidence has been assembled to show that pay-for-performance incentives are either widely used in public agencies or that they are effective when used. The exercises we offer here provide two possible explanations for this state of affairs: that in the first case, public agents face such low efficacy that the construction of an incentives system is irrational; that in the second case, public principals face a state of affairs where they find it not in their own self-interest to impose incentives-based systems even though agents would respond to them in the socially desired way.

Congress only recently moved to allow agencies pay significantly greater salaries to executives who work under rigorous performance-ranking systems. The Office of Personnel Management now struggles with certifying such systems. Historically, such ranking systems have consistently ranked executives in the top tier: in fiscal year 2000, 85.4% of career SES members at the Environmental Protection Agency rated its highest level; 61.4% of those at the Treasury Department rated at the highest level (Barr 2004). Those numbers have dropped as greater scrutiny is placed in performance-ranking systems (in 2002, to 48.7% at Treasury and 61.4% at the EPA). Monetary incentives, though, are only weakly tied to performance rankings: at Health and Human Services, only 37.1% of its highly rated

Miller and Whitford  The Principal’s Moral Hazard 225
executives were given awards in 2002 and the average bonus was only $10,307. There is no such provision for bonus payments for those in the traditional civil service in the United States.

NEGATIVE BONUSES AND WAGES

Low levels of individual efficacy require extremely high rewards. But there is a way around this problem if the legal system will allow them. The solution is to use negative incentives combined with a flat wage. For instance, the agent would receive a healthy flat wage that is not contingent on the outcome but would be fined an amount, Z, in the event of failure. Negative incentives are the contractual equivalent of sanctioning behavior. We recognize that principals (especially political ones) may try to use ex post sanctions to penalize agents (especially public personnel) in work settings. We are specifically concerned with individual-level sanctions and not group-level sanctions like reduced budgetary resources, legislation that limits policy or administrative discretion, or other ways of manipulating an agency’s jurisdiction. We are most concerned about individual-level sanctions because this is the domain of incentives-based approaches to enhanced performance and group-level sanctions presumably have individual-level consequences. We further address group-level monitoring and incentives contracts below.

Mathematically, the amount of the fine must be of the same magnitude as the bonus: $Z = -B$. That is, with the illustrative asset and cost parameters, the agent must receive a fine of at least $16,000. The agent would supply a HI effort to reduce the probability of the $16,000 fine rather than to increase the probability of the $16,000 bonus. When $p = 0.8$, the agent will have a 20% chance of paying the fine even if with a HI effort. Naturally, the agent must receive a fixed wage $W$ to compensate for both the effort cost and the possibility of the fine. This wage $W$ must be

$$(1 - p)B^* + C_{HI} = $11,200.$$

This leaves the agent with none of the expected social surplus, whereas the principal retains all the social surplus. In other words, the possibility of a negative fine results in the principal’s expected profits being equal to the entire expected social surplus. There is no longer a wedge between the principal’s self-interest and efficiency.

So as long as the agent is risk neutral, a negative fine eliminates the principal’s moral hazard. The problem is that the expected net gain of zero is the result of two possibilities: either the outcome is a success and the agent clears $3,200 net gain or the outcome is a failure and the agent owes the principal $16,000. The latter possibility violates what might be called a bankruptcy constraint: will the agent have savings sufficient to cover the fine? Can the city charge the police officer an amount she may not have in the event of a crime wave that was beyond her control? If not, then the negative bonus plan is not going to solve the problem of the principal’s moral hazard. Moreover, this assumes risk

---

4 The National Defense Authorization Act for fiscal year 2004 (Public Law 108-136, November 24, 2003) created a performance-based pay system for members of the SES. Under that system, executives in an agency with a Certified SES Performance Appraisal System can be paid a minimum flat wage of $103,700 and a maximum of $157,000. Those in agencies without a Certified SES Performance Appraisal System can be paid a minimum flat wage of $103,700 and a maximum of $144,600.
neutrality; in the more usual case, a risk-averse agent would not accept a negative bonus incentive plan.

The same thing is true if the principal can charge a negative flat wage. The principal might try to guarantee she earns the entire social surplus when paying a bonus of $B^*$ by making the compensation package be a combination of the bonus $16,000 plus a negative flat wage of $4,800. In other words, the agent buys partial ownership for $4,800 and receives the bonus in case of success. With $W = -4,800$, then the principal’s expected profit when paying a bonus of $16,000 is exactly equal to the social surplus: $18,000.

To revert to the police patrol example: imagine that the patrol officer has to pay a weekly amount for the right to work as a police officer. With a sufficiently large payment, the city could afford to offer a sufficiently large bonus for crime reduction. The crime fighter would pay the up-front fee for the right to earn the sufficiently attractive bonus. This system would have the added advantage of solving an adverse selection problem. The crime fighters who were least confident of their abilities to reduce crime would not pay the up-front fee, whereas the crime fighters who were most confident of their abilities to reduce crime (and thereby earn the large bonus) would be the ones who would be most willing to seek out the job. Note that a large up-front fee may also cause other departments to suffer from adverse selection problems as less-able cops flow to jurisdictions with guaranteed contracts.

Selling an ownership share in the asset to the agent is a standard solution in principal-agency theory (e.g., Dixit and Nalebuff 1993). The beauty of this solution is that it reconciles efficiency with the principal’s self-interest by guaranteeing that the principal receives the entire social surplus.

Yet, in practical terms, selling “ownership” is restricted for public employees. Public employees may not be allowed or required to have an ownership share in the organizations in which they are employed. If ownership is shares held by citizens in a democracy, no person holds sufficient shares to effectively bind that individual’s discretionary choices (as in rational voting abstention). Likewise, the penalization of agents is difficult in many contractual settings, but especially so in government. In particular, a negative wage or fee for accepting a job will also violate a bankruptcy constraint as the employee could pay the up-front fee in hopes of the sizable bonus, only to have (by the luck of the draw) a failure and no bonus.

**INDIVIDUAL EFFICACY IN TEAMS**

Our claim in this article is that the principal contracting with a single agent will underutilize incentives because they detract from their own profits even when they contribute to efficiency. In this section, we argue that the same thing is true when principals contract with teams of agents—there is a wedge between efficiency and the principal’s profits.

---

5 We do not want to overdo the distinction between public agencies and firms, although there is evidence that the use of ownership schemes has increased in the private sector in recent decades (Conte and Kruse 1991; Jones and Kato 1995; Knez and Simester 2001). Blasi, Conte, and Kruse (1996) find moderate performance differences for firms with and without employee ownership, and additional empirical support is found in studies of self-employed contractors (Andersen and Schmittlein 1984) and franchises (Lafontaine 1992; Slade 1996). Estimates of the productivity improvements from the use of employee stock option plans often run to 4%–5%. But our general point is the more important one: if firms do not use employee ownership schemes it is usually due to choice, whereas in government, ownership is restricted by constitution.
Some technologies of production rely on teams of agents, in which each agent’s individual contribution to overall team production levels may be low. This creates an additional incentive for shirking—an individual member of the police department may rationalize that his own frequent trips to the doughnut shop have a negligible effect on the department’s ability to provide public safety. Here again, we start with a problem of agent moral hazard, but we end up concluding that the principal’s moral hazard is the ultimate problem.

To illustrate this, imagine a team of \( n \) individuals producing a public good, such as a road, an irrigation system, or a crime-control program. (The team members could include ordinary citizens, as in a Neighborhood Watch program for crime control.) The public good, if delivered, results in real economic benefits. Is there any way to allocate the benefits of team production so that each team member has an incentive to work effectively? In particular, we would like the allocation scheme to be budget balancing—that it allocates the benefits so that the benefits allocated to the team collectively equal exactly the benefits produced. Holmstrom (1982) has shown that there is no such scheme. Every individual team member has an incentive to shirk in recognition that it makes only a slight difference to the team member’s own share of the benefits.

Holmstrom’s own solution to this dilemma was very much in the spirit of our own essay—to create a mechanism for making each member of the team feel efficacious, through what he called a “joint-forcing contract.” The joint-forcing contract specified that each team member would receive compensation covering his effort cost if and only if the efficient production level is achieved. Any one individual’s shirking then deprives everyone (including himself) of his compensation. This hair-trigger punishment creates a Nash equilibrium in which every team member finds it rational to work efficiently.

With this contract, the nonproductive “residual owner” becomes the “budget breaker,” allowing a reconciliation of efficiency with self-interest. Each team member knows that his own shirking would deprive himself (along with everyone else) of compensation. Thus, efficient levels are a Nash equilibrium, no matter how tiny each team member’s contribution is to the overall aggregate team production. The purpose of the joint-forcing contract is to create a kind of “artificial” (good) efficacy for each individual, a knife-edge contract that makes each team member’s contribution decisive for everyone’s compensation.

The joint-forcing contract requires a “ruler.” The ruler is required, first of all, to write and enforce the joint-forcing contract. Furthermore, the ruler is required to receive the residual benefits after paying off (minimally) the team members. The ruler feeds and clothes the workers and then appropriates all the remaining wealth for himself (e.g., pyramids, official art, the army).

Holmstrom’s is a clever result that, in one sense, gets around a portion of the result from earlier in this essay. High levels of individual efficacy are not necessary to support the use of incentive contracts. The joint-forcing contract renders unnecessary any effort to...
monitor, direct, or sanction subordinate behavior. The principal (budget breaker) can rely on the agent’s self-interest alone to support the efficient Nash equilibrium.

Yet the principal’s own self-interest is a problem. By instituting this contract, the principal could support an efficient Nash equilibrium driven by nothing but the agent’s self-interest; but the principal’s own self-interest is served by subverting the incentives contract.

This point is made in Eswaran and Kotwal (1984), who demonstrate that the principal has every incentive to pay one team member to shirk. This side payment makes it unnecessary to pay any of the team members, due to the knife-edge result of the joint-forcing contract. The principal is laden with her own form of moral hazard. The other team members must know the temptation facing the principal; as a result, they will expect the subversive side payment. And in expecting this outcome, they will collapse to a Nash equilibrium in which no one works at all. The ruler’s temptation to “cheat” on a joint-forcing contract subverts overall efficiency.

The irony is that the creation of a kind of artificial individual efficacy under the joint-forcing contract cuts both ways—it means that one agent is just as efficacious in subverting the efficient contract as she is essential to promoting it. Combined with the moral hazard of incentive setting, this means that once again, self-interest by itself is not sufficient to guarantee the irrelevance of command and coercion. Moreover, whereas Eswaran and Kotwal (1984) demonstrate this in the case of team production, we demonstrate this even in the canonical agency problem, where one principal contracts with one agent, and in the simplest possible case, where both the agent’s individual rationality constraint is satisfied and social surplus is maximized.

**USING INCENTIVES REQUIRES CONSTRAINING SELF-INTEREST**

Eswaren and Kotwal use the Holmstrom impossibility result to demonstrate the opposite of what Holmstrom originally intended: rather than showing that self-interest and efficiency are reconcilable in the absence of budget balancing, budget breaking relocates moral hazard from team members to the budget breaker. Once again, one of the reasons we will not see incentives used in organization is not that they cannot work but that the self-interest of principals prevents their use. A precondition for using incentives in organizations is, ironically, the credible constraint of the principal’s self-interest. A principal who is unable to derail the efficient incentive scheme can induce the efficient Nash equilibria.

In fact, a primary form of constraint may well be reputation. In a repeated game, the residual owner could realize benefits from establishing a reputation for integrity—by not subverting the efficient incentive scheme. The team members can maintain the belief that the residual owner will not subvert the process in a shared recognition that this would result in the imposition of a credible punishment strategy (in the form of general reversion to a shirking equilibrium). This seems to be the case with Lincoln Electric (Miller 1992, 116). Yet, counterexamples from recent business news such as Tyco, Enron, and Adelphia show how leaders may choose to subvert the process, with attendant consequences.

Alternatively, the selection of agents becomes more important when outcome-based incentives cannot align self-interest and organizational efficiency. People with a strong sense of public cause may be more important to bureaucracy for that reason. Intrinsic motivation (or valence) from the performance of work tasks is widely recognized in work settings (Vroom 1964) and occurs in many public agency settings.
We note, however, that recent studies have revisited the question of selection of individuals with a strong sense of mission in public settings and the binding role of belief systems in constraining public employees in the face of moral hazard (e.g., Brehm and Gates 1999). This, of course, has been long debated in public administration, with a common argument being that administrators are bound by a strong sense of professionalism (Mosher 1968). This perspective relies on a model of the cognitive process where values bind individual decisions, a strong component of the view of March and Simon (1958). Evidence exists that public administrators do have distinct value preferences from other preferences (i.e., law, business) that may instill a sense of extraordinary ownership in the public product (e.g., Nalbandian and Edwards 1983). We revisit the effects of public service motivation (PSM) here.

Essentially, theories of PSM argue that public personnel are imbued with an ethic to serve the common good by working in public agencies. Most importantly, PSM characterizes traditional aspects of organizational behavior in public agencies like performance, effectiveness, and the sorting of personnel (e.g., Rainey 1982; Perry and Wise 1990; Romzek 1990; Crewson 1997; Brewer and Selden 1998). More broadly, PSM can help explain voluntary behavior and blood donations (Houston 2005) and counterbalances to declining social capital (Brewer 2003). Brehm and Gates (1999) document this to a strong degree, and professionalization of the public sector only enhances this ability. To some extent, this is what W. Wilson (1987) argued in his famous claim for the “politics-administration dichotomy,” which was much more a prescriptive offering on the state of the public sector at the turn of the century than a positive statement about the constitutional role of the public sector in a system of shared and separated powers.

Essentially, our argument provides one more rationale for the enhanced study of solutions to the adverse selection problem, rather than a continued concentration on the principal’s solution to the canonical moral hazard problem. However, we want to be clear that the solution of adverse selection problems is different from solving the problem of malformed incentives of the principal. If a principal desires successful outcomes, our result does not address how they may go about attaining that outcome. Rather, our result makes clear that principals may—in some cases—desire unsuccessful outcomes that actually reduce social surplus (but leave the principal with a larger share). Our suspicion is that solving the adverse selection problem in these specific cases may simply shift a greater share of the outcome to the principal—that the principal will obtain successful outcomes without having to compensate the agent for providing the high effort that produces success. In this sense, institutional constraints that require principals to select workers of high quality will produce greater social surplus; what remains to be seen is how those rules apportion that surplus (how much the principal gets to keep). We also reiterate that any conclusions about how solving the adverse selection problem affects the motivations for principals that we have laid out here are largely speculative since the adverse selection and moral hazard problems are largely separated in theoretical development.

In fact, our general position—that is, for entire classes of production problems, canonical principal-agency prescriptions about incentives are insufficient to cause agents to “do the right thing”—is reinforced by a theory of democracy in which separated powers define institutional relations. Agencies are called to produce when agent efficacy is especially low. In these cases, the problem for political officials is to select appropriate agents and create opportunities for them to seek the public interest—even though their individual efficacy says they should not. Separated powers may make it possible for such agents to
flourish because such principals are less likely to coordinate their behavior and constrain the efforts of agents trained to seek the public product. This is particularly the case when incentivization is not rational for principals—when political actors may not individually see value in causing agents to offer higher effort levels. In that case, well-minded agents act as checks on the power of the principal to subvert the social good. In this sense, the type of agent that makes the principal’s job easier when agent efficacy is especially low—the agent that produces higher effort without incentives—is the same type that makes society’s job easier when the principal experiences his own type of moral hazard (produces higher effort when the principal does not desire it). Democratic performance is contingent on agents because they are employed exactly when incentives often fail, either out of inefficacy or out of subversion.

REFERENCES


