# Optimal Incentive Contract with Costly and Flexible Monitoring

Anqi Li* Ming Yang†

September 2016

**Abstract**

Recent advances in IT and big data enable firms to adopt an increasing variety of monitoring technologies at a reduced and yet significant cost. We examine the effect of such cost and flexibility on employee productivity and the internal organization of firms. In an otherwise standard principal-agent model with moral hazard, we allow the principal to adopt any monitoring technology that constitutes a finite partition of the agent's performance state space, at a cost that increases as the induced performance measure becomes more fine-grained. In various classical settings, we examine the optimal incentive contract through trade-off between the compensation cost and the monitoring cost, obtaining characterizations, such as information aggregation, strict MLRP, the fine-tuning of monitoring intensity across tasks according to the agent's tendency to shirk, and the use of group incentive systems among technologically independent agents. We apply these results to human resource management and suggest new explanations for long-lasting puzzles.

*Department of Economics, Washington University in St. Louis. anqili@wustl.edu.
†Fuqua School of Business, Duke University. ming.yang@duke.edu.

# 1 Introduction

Recent advances in IT and big data bring new opportunities and challenges to employee monitoring. First, a growing volume and variety of performance information can be processed, stored and reported at a reduced and yet significant cost,[1] enabling all-round performance appraisals based on the 360-degree feedback received from an employee's supervisors, peers, subordinates and customers (Bracken et al. (2001)), as well as detailed records that tracks his time-spending patterns and communication histories (Woodley (2013), Straz (2015)). Second, more flexibility is embedded in the design and implementation of monitoring technologies, sparking discussions about how we can classify the various sources of feedback into meaningful categories (Pulakos (2004)), which rating scale balances informativeness and interpretability (Hook et al. (2011)), how managers should trade off the monitoring of linked activities (Kaplan and Norton (1992, 1993)), and what mix of individual and group incentives can best motivate a large group of employees (Bryson et al. (2013)).

As of today, multi-source feedback tools are widely adopted across the globe,[2] and big data analysis is gaining momentum in human resource management.[3] These trends raise an important question, that of how the cost and flexibility they introduce to can potentially affect the monitoring and rewarding of employees, which in turn has important effects on employee productivity and the internal organization of firms. However, the existing incentive theory is ill-suited for addressing this question, as most models we have seen either ignore the monitoring cost or severely limit firms' choices over monitoring technologies. The current paper takes a step towards filling this gap.

Our framework builds on an otherwise standard principal-agent model with moral hazard, where we represent all acquirable information about the agent's hidden effort by a random and potentially high-dimensional performance state. Motivated by real-

[1] On the gain side, Ewen and Edwards (2001) estimates that web-based technologies have reduced the administration cost of multi-source feedback by as much as 80 percent; Baker and Hubbard (2004) studies how on-board computers enable the better monitoring of truck drivers; and Solman (2013) reports that cloud-based technologies are increasingly used for employee performance tracking and analysis. On the cost side, Bracken et al. (2001) details how the infusion of new data complexifies the implementation of multi-source feedback; a survey by Towers and Watson in 2014 ranks HR data and analytics among the top three areas for HR technology spending; and a recent CB Insights article reports how big data has spurred the growth and M&A of HR startups ("The Data-ification of HR," 2015).

[2] This includes at least one third of U.S. companies and 90 percent of Fortune 500 companies.

[3] "The Big Data Opportunity for HR and Finance," *Harvard Business Review*, April 24, 2016.

world practices, we allow the principal to adopt any monitoring technology that constitutes a finite partition of the performance state space, at a cost that increases as the induced performance measure becomes more fine-grained. Assuming that the agent can only be compensated based on the realization of the *monitoring outcome*, i.e., the cell of the partition that contains the realized performance state, we examine the optimal incentive contract through the trade-off between the compensation cost and the monitoring cost, obtaining characterizations, such as information aggregation, the strict monotone likelihood ratio property, the fine-tuning of monitoring intensity across tasks according to the agency's tendency to shirk, and the use of group incentive systems among technologically independent agents. We apply these results to human resource management and suggest new explanations for long-lasting puzzles.

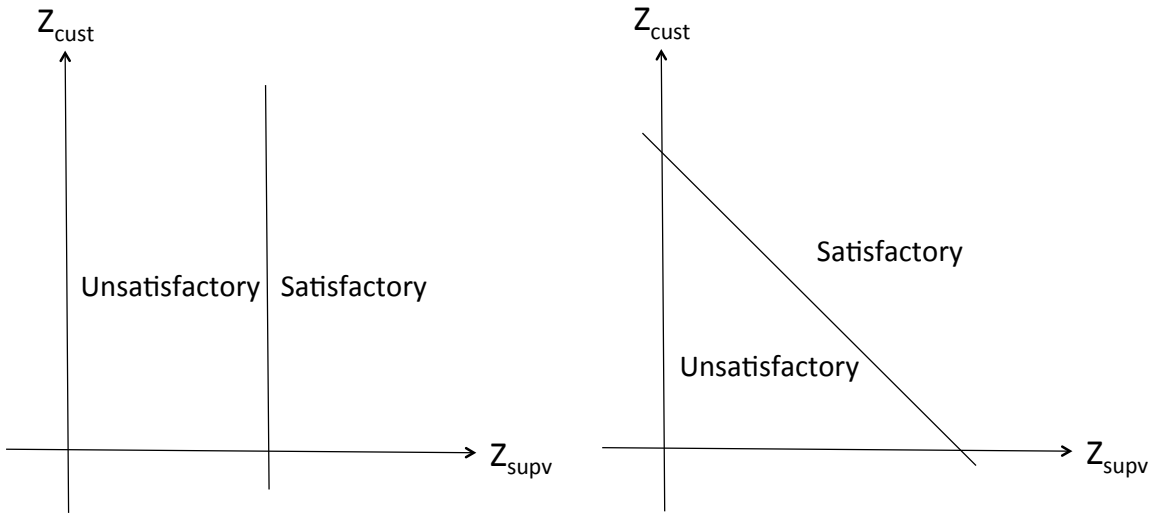


Figure 1: Commonly used monitoring technologies represented by partitions of the performance state space.

To illustrate our framework, suppose the agent's performance state consists of a supervisory evaluation and a customer review. In a hypothetical world where monitoring is costless, fully revealing every performance state provides the agent with the strongest incentive to work. But in reality, the processing, storage and communication of performance data incur significant costs, giving rise to a common practice that is supported by information theory (Cover and Thomas (2006)), that of classifying the fine-grained performance data into a limited number of categories (Bracken et al. (2001), Pulakos (2004), Hook et al. (2011)). Figure 1 depicts two monitoring

technologies that achieve this goal, where the one on the left panel divides the performance states into "*satisfactory*" and "*unsatisfactory*" depending on whether the supervisory evaluation exceeds a cutoff or not, whereas the one on the right panel does so according to whether a weighted score is above or below a threshold. Admittedly, these monitoring technologies put emphasis on distinct information sources and assign varying contents to the monitoring outcomes. But in the current framework, they can both be adopted by the principal, as long as the performance measures they induce have about the same degree of fine-grainedness. Together, these assumptions formalize the cost and flexibility that arise in the design and implementation of monitoring technologies, enabling analysis of their impact on employee productivity and the internal organization of firms.

The optimal monitoring technology balances the trade-off between the compensation cost and the monitoring cost. To illustrate this idea, we first revisit the classical setup of Holmstrom (1979), where a single agent can influence the distribution of performance states by exerting either the high effort or the low effort. In this setting, the optimal monitoring technology features *information aggregation based on a cutoff rule over likelihood ratios*, meaning that it compresses the fine-grained and high-dimensional performance states with similar likelihood ratios into the same coarse, single-dimensional grades. This result explains why multi-source performance appraisal systems aggregate the various sources of feedback into coarse, rank-ordered ratings, such as "outstanding," "highly effective," "satisfactory" and "unsatisfactory," according to the employee's performance measured by an overall score. It also justifies the assignment of coarse grades, such as A, B, C, D and F, based on the student's performance in terms of an overall grade.

This result showcases Holmstrom (1979)'s *sufficient statistics principle*, which says that the optimal wage scheme for any given monitoring technology depends only on the likelihood ratio of the monitoring outcome. This suggests that when monitoring is costly and flexible, the principal should focus on the processing, storage and communication of the likelihood ratio, and ignore the part of performance state that is orthogonal to the likelihood ratio. As a result, the optimal monitoring technology assigns distinct average likelihood ratios to different monitoring outcomes, and therefore satisfies the *strict monotone likelihood ratio property* with respect to the order over likelihood ratios (henceforth abbreviated as strict MLRP). In addition, since performance states with similar likelihood ratios have similar effects on the compensation

cost, whereas monitoring is flexible, meaning that the monitoring cost is independent of the monitoring outcomes' likelihood ratios, it follows that the optimal monitoring technology classifies performance states with similar likelihood ratios into the same monitoring outcome and hence can be obtained from applying a simple cutoff rule to the space of likelihood ratios. These results remain valid even if we allow for random monitoring technologies or if part of the performance state (e.g., company's cash flow) can be observed at no cost. Drawing on the findings of Bloom and Van Reenen (2006, 2007, 2010), they attribute the use of different monitoring technologies to factors that affect the (opportunity) cost of distinguishing employee performance (e.g., access to IT, labor market regulation, product market competition), thereby adding a new explanation to the long-lasting puzzle surveyed by Gibbons and Henderson (2012), that of why the management practices adopted by otherwise similar firms exhibit significant and persistent heterogeneity.

In case the agent can take multiple deviant actions, the optimal incentive contract takes the form of a *balanced scorecard* (Kaplan and Norton (1992, 1993)), whereby the resources spent on the detection of each potential deviation match the Lagrange multiplier of the corresponding incentive compatibility constraint. In the multi-task model considered by Holmstrom and Milgrom (1991), this suggests that the principal fine-tune the monitoring intensity across tasks according to the agent's tendency to shirk. This result has policy implications, including how universities should evaluate the teaching performance of faculties who simultaneously engage in other tasks, such as research and administration.

We finally turn to the multi-agent model considered by Holmstrom (1982), Green and Stokey (1983) and Mookherjee (1984), where the conventional wisdom attributes the use of group incentive contracts (e.g., team and tournament) to either the effort complementarity or the common productivity shock between agents on the one hand, and limits the use of individual incentive contracts among technologically independent agents on the other hand. Recently, this view has been challenged by Bloom and Van Reenen (2006, 2007), who find that even firms with similar production technologies make significantly different choices between individual and group incentive contracts. We resolve this puzzle from the angle of monitoring cost. Intuitively, group incentive contracts lump the assessment of agents together and yield coarser performance ratings than individual incentive contracts. When monitoring is costly and flexible, the limited monitoring capacity creates an *attentional linkage* between agents and

stipulates the use of group incentive contracts even if agents are technologically independent. The main prediction of our result, that employees are less recognized for their individual performance as the monitoring cost increases, other things equal, is supported by the findings of Bloom and Van Reenen (2006, 2007).

## 1.1 Related Literature

Most foundational works on incentive contracting take the monitoring technology as exogenously given, including the single-agent model of Holmstrom (1979), the multi-task model of Holmstrom and Milgrom (1991) and the multi-agent models of Holmstrom (1982), Green and Stokey (1983), Nalebuff and Stiglitz (1983) and Mookherjee (1984). Meanwhile, existing studies on contracting with costly monitoring impose strong limitations on the principal's choice over monitoring technologies. For example, in the costly verification model developed by Banker and Datar (1980) and Dye (1986), the principal is limited to drawing a signal from an exogenously given probability distribution. And in the linear contracting model that appears commonly in applied works, the principal can only pay to reduce the variance of a Gaussain performance signal. Due to the lack of flexibility in the choice over monitoring technologies, these models cannot jointly predict as our model does: information aggregation, strict MLRP, the fine-tuning of monitoring intensity across tasks, and the use of group incentive contract among technologically independent agents.

A growing body of empirical studies has examined the impact of IT on the internal oragnization of firms. In particular, Milgrom and Roberts (1992) documents how the price decline of IT accelerated the replacement of mass production with modern manufacturing; Caroli and Van Reenen (2001) investigates the role of IT in facilitating decentralization and multi-tasking; Bresnahan et al. (2002) exploits the complementarity between IT, workforce skill and organization changes such as employee autonomy and teamworking; and Bloom et al. (2012) attributes the U.S.'s IT related productivity advantage to its tough people management policies on promotions, rewards, hiring and firing. In constrast, we formalize the cost and flexibility that IT brings to employee monitoring and characterize the optimal monitoring technology in various classical settings.

The current paper shares a common spirit with the literature on rational inattention pioneered by Sims (2002) and Sims (2006), whereby economic agents can process

the information required for decision making in a flexible manner, subject to a limited capacity constraint measured by Shannon entropy. Recently, this framework has been applied to the study of various macro and microeconomic probelms. For example, Maćkowiak and Wiederholt (2009) explains observed patterns on price stickiness by the optimal allocation of firm's attention between idiosyncratic and aggregate shocks. Matéjka and McKay (2012) examine the optimal pricing strategy against consumers who can process the offers made by firms with a certain degree of flexibility. The current paper differs from these studies in two respects: first, we focus mostly on partitional monitoring technologies in order to best represent reality, though our results carry over qualitatively to random monitoring technologies; second, we consider a large class of monitoring cost functions that nests entropy as a special case.

Several authors have examined the impact of limited communication on mechanism and organization design. Among them are Dye (1985), which characterizes the outcome of bilateral trade when state contingencies are costly to write; Blumrosen et al. (2007), which derives the optimal auction format in case bidders can send only a few messages to the auctioneer; Crémer et al. (2007) and Sobel (2015), which characterize the optimal organizational code that can be described by a finite vocabulary; Dessein et al. (2016), which examines the trade-off between coordination and localization when the communication between team members is costly; and Green and Laffont (1987) and Madarasz and Prat (2016), which investigate screening problems where the agent cannot fully describe his hidden type.

A vast literature is devoted to understanding why heterogeneity prevails among the management practices adopted by otherwise similar firms (see Gibbons and Henderson (2012) for a thorough survey). Recent theoretical works on this subject matter include but are not limited to Chassang (2010), Li and Matouschek (2013) and Halac and Prat (2014). Specifically, Chassang (2010) formalizes the adoption of new monitoring technology as a bandit problem and obtains characterizations, such as heterogeneity and path-dependence. Li and Matouschek (2013) examines the conflict cycle in relational contracts where a random and private shock affects the principal's cost of paying bonuses. And Halac and Prat (2014) characterizes the equilibrium monitoring intensity in a reputation model, where the market imperfectly observes a firm's decision on whether or not to monitor its employee.

The remainder of this paper proceeds as follows: Section 2 introduces the model setup; Section 3 presents the main results; Sections 4 and 5 investigate extensions of

the baseline model; Section 6 concludes. See Appendix A for omitted proofs and the online appendix for further results.

# 2 Setup

**Players** There is a risk-neutral principal (she) and a risk-averse agent (he). The agent can spend a non-negative wage $w \geq 0$ and privately exert either the high effort ($a = 1$) or the low effort ($a = 0$). His utility is given by $u(w) - c(a)$, where $u(0) = 0$, $u' > 0$, $u'' < 0$ and $c(1) = c > c(0) = 0$. Each effort $a$ generates a probability space $(\Omega, \Sigma, P_a)$, where $\Omega \subset \mathbb{R}^d$ is the *performance state space*, $\Sigma$ is the Borel sigma-algebra restricted to $\Omega$, and $P_a$ is the probability measure over $(\Omega, \Sigma)$ given $a$. In particular, each *performance state* $\omega \in \Omega$ contains all acquirable information about the agent's hidden effort (e.g., 360-degree feedback), whereas $P_a$ is equipped with a well-defined probability density function $p_a$. The principal's *expected* payoff depends only on the agent's effort. Her goal is to induce the high effort through the use of incentive contracts.

**Incentive contract** An incentive contract $\langle \mathcal{P}, w(\cdot) \rangle$ is a pair of *monitoring technology* $\mathcal{P}$ and *wage scheme* $w : \mathcal{P} \to \mathbb{R}_+$. In general, a monitoring technology is a probabilistic mapping between the performance state space and finitely many monitoring outcomes. To highlight the main intuition, we focus on *partitional monitoring technologies* in the main body of this paper and defer the discussion about general monitoring technologies to Appendix B.4. Specifically, let $\mathcal{P}$ be any finite partition of $\Omega$ whose cells belong to $\Sigma$, and $w : \mathcal{P} \to \mathbb{R}_+$ be a function that maps each cell $A$ of $\mathcal{P}$ to a non-negative wage $w(A) \geq 0$. For each $\omega \in \Omega$, use $A(\omega)$ to denote the unique *monitoring outcome* that contains $\omega$, and let $w(A(\omega))$ be the wage payment at state $\omega$. Time evolves as follows:

1. The principal commits to an incentive contract $\langle \mathcal{P}, w(\cdot) \rangle$;
2. The agent privately exerts an effort $a \in \mathcal{A} = \{0, 1\}$;
3. Nature draws a performance state $\omega \in \Omega$ according to $P_a$;
4. The monitoring technology publicly announces the monitoring outcome $A(\omega)$;
5. The principal pays the promised wage $w(A(\omega))$ to the agent.

Each pair of monitoring technology $\mathcal{P} = \{A_1, \cdots, A_N\}$ and effort $a$ defines a random *performance measure* $X : \Omega \to \mathcal{P}$ whose p.m.f. $P_X(\cdot \mid a)$ is given by $P_X(X = A_n \mid a) = P_a(\omega \in A_n)$ for all $A_n \in \mathcal{P}$. Let $\vec{\pi}(\mathcal{P}, a) = (P_X(X = A_1 \mid a), \cdots, P_X(X = A_N \mid a))$ denote the probabilities of the performance measure that $(\mathcal{P}, a)$ induces.

**Implementation cost** At any given level $a$ of agent effort, the total cost of implementing an incentive contract is given by

$$\sum_{A \in \mathcal{P}} P_a(A) w(A) + \mu \cdot H(\mathcal{P}, a).$$

This cost has two parts. The first part $\sum_{A \in \mathcal{P}} P_a(A) w(A)$, or the *compensation cost*, has been the central focus of the existing principal-agent literature. Meanwhile, the second part $\mu \cdot H(\mathcal{P}, a)$, henceforth referred to as the *monitoring cost*, is new and captures the cost that is associated with the processing, storage and communication of performance data. Throughout, we assume that the monitoring cost equals the product of (1) $H(\mathcal{P}, a)$, a measure of the fine-grainedness of the random performance measure that $(\mathcal{P}, a)$ induces, and (2) $\mu > 0$, an exogenous variable that parameterizes the difficulty in implementing fine-grained monitoring technologies (henceforth referred to as the *marginal monitoring cost*).

Inspired by information theory, we make the following assumption on the monitoring cost function.

**Assumption 1.** *There exists a function $h$ such that $H(\mathcal{P}, a) = h(\vec{\pi}(\mathcal{P}, a))$ for all $(\mathcal{P}, a)$. For any $N \in \mathbb{N}$ and $(\pi_1, \cdots, \pi_N) \in \Delta^N$,*

*(a) $h(\pi_1, \cdots, \pi_N) = h\left(\pi_{\Pi(1)}, \cdots, \pi_{\Pi(N)}\right)$ for all permutation $\Pi$ over $\{1, \cdots, N\}$;*

*(b) $h(\pi_1, \cdots, \pi_N) < h(\pi_1', \pi_1'', \cdots, \pi_N)$ for all $\pi_1', \pi_1'' > 0$ such that $\pi_1' + \pi_1'' = \pi_1$.*

Assumption 1 says that the monitoring cost is invariant to how we assign contents or namings to the monitoring outcomes, and that it increases as the induced performance measure becomes more fine-grained. This assumption plays a crucial role in formalizing the cost and flexibility that are arise in the design and implementation of monitoring technologies. In information theory, it is satisfied by many commonly used measures for the quantity of information, including bits and entropy (Cover and

Thomas (2006)). Due to space limitations, we will further explain this assumption and give examples of the fine-grainedness measure in Section 2.1. We will defer the discussion on marginal monitoring cost to Section 3.3.

**$z$-value** Suppose the *likelihood ratio* $p_0/p_1$ exists, and define a random variable $Z : \Omega \to \mathbb{R}$ by

$$Z = 1 - \frac{p_0}{p_1}.$$

For each $A \in \Sigma$, use $Z(A)$ to denote the image of $A$ under the mapping $Z$, and define the *$z$-value* of $A$ by

$$z(A) = \mathbb{E}[Z \mid A; a = 1].$$

A contract is *incentive compatible* if it induces the agent to exert the high effort, i.e.,

$$\sum_{A \in \mathcal{P}} u(w(A)) P_1(A) z(A) \geq c. \tag{IC}$$

A close inspection of the (IC) constraint reveals that $z$-value contains all the information that the principal needs in order to deter shirking.

**Optimal incentive contract** A contract satisfies the agent's limited liability constraint if (the online appendix replaces this with an individual rationality constraint)

$$w(A) \geq 0, \forall A \in \mathcal{P}. \tag{LL}$$

An optimal incentive contract that induces the high effort from the agent minimizes the total implementation cost, subject to the agent's incentive compatibility constraint and limited liability constraint, i.e.,

$$\min_{\langle \mathcal{P}, w(\cdot) \rangle : |\mathcal{P}| \in \mathbb{N}} \sum_{A \in \mathcal{P}} P_1(A) w(A) + \mu \cdot H(\mathcal{P}, 1), \text{ s.t. (IC) and (LL).} \tag{2.1}$$

In Appendix B.1, we show that the solution to $\langle \mathcal{P}^*, w^*(\cdot) \rangle$ to this problem exists under mild regularity conditions. One can verify that these conditions are compatible with the assumptions that are required for establishing our main results. Unless otherwise specified, our statements hold true except perhaps on a measure-zero set of states.

## 2.1 Illustrative Examples

This section illustrates the usefulness of our framework through examples. We begin with the observation that partitional monitoring technologies are commonly used for classifying performance data in human resource management.

**Example 1.** Suppose $\Omega = \mathbb{R}$ and each $\omega \in \Omega$ represents an outcome of supervisory evaluation. When monitoring is costless, fully revealing every $\omega$ gives the agent the strongest incentive to exert high effort. But in reality, the processing, storage and communication of performance data incur significant costs, giving rise to the common practice of classifying the fine-grained performance data into coarse categories. An example would be to label those evaluations that lie above a threshold $\widehat{\omega}$ as "*satisfactory*" and those that fall below $\widehat{\omega}$ as "*unsatisfactory.*" This is formally achieved by implementing the monitoring technology $\mathcal{P} = \{(-\infty, \widehat{\omega}), [\widehat{\omega}, +\infty)\}$ that partitions $\Omega$ into $(-\infty, \widehat{\omega})$ and $[\widehat{\omega}, +\infty)$.

We next explain how our assumption — that any partitional monitoring partition can be implemented at a cost that increases with the fine-grainedness of the induced performance measure — helps formalize the flexibility that arises in the design and implementation of monitoring technologies.

**Example 2.** Suppose $\Omega = \mathbb{R}^2$ and each $\omega \in \Omega$ consists of a supervisory evaluation $z_{supv}$ and a customer review $z_{cust}$. Figure 1 depicts two commonly observed monitoring technologies, where the one on the left panel focuses exclusively on the "*downward feedback*" that is given by the supervisor, whereas the one on the right panel takes into account the "*upward feedback*" that is received from the customer. Admittedly, these monitoring technologies put emphasis on different information sources and assign varying contents to the monitoring outcomes "*satisfactory*" and "*unsatisfactory.*" But under Assumption 1, they incur the exact same monitoring cost, as long as the performance measures they induce have the same degree of fine-grainedness measured by the probabilities of monitoring outcomes.

To better understand this last assumption, suppose, to the contrary, that the monitoring cost can depend on the contents of the monitoring outcomes, too. Then in general, the above described monitoring technologies incur different monitoring costs even if they have the exact same degree of fine-grainedness measured by the probabilities of monitoring outcomes. In this sense, we have created an artificial barrier to the

adoption of the expensive monitoring technology and limited the principal's choice for reasons beyond the cost of processing, storage and communication of performance data. Assumption 1 removes this barrier and expands the principal's choice set over monitoring technologies.

We finally give a commonly used fine-grainedness measure that satisfies Assumption 1.

**Example 3.** As noted by Bracken et al. (2001), Pulakos (2004) and Hook et al. (2011),[4] the number of performance categories, often termed as the "*rating scale*," is commonly used to measure the fine-grainedness of the performance appraisal system in human resource management. Monitoring cost functions that fit this description take the form of $f(|\mathcal{P}|)$, where $|\mathcal{P}|$ denotes the cardinality of $\mathcal{P}$ and $f$ is an increasing function over $\mathbb{N}$.

# 3 Main Result

This section examines the main features of the optimal monitoring technology. Specifically, Section 3.1 shows that the performance measure induced by the optimal monitoring technology satisfies the *strict monotone likelihood ratio property* with respect to the order over $z$-values. Meanwhile, Section 3.2 demonstrates that the optimal monitoring technology achieves *information aggregation based on a cutoff rule over $z$-values*, meaning that it groups fine-grained and high-dimensional performance states with similar $z$-values to the same coarse, single-dimensional grades.

## 3.1 Strict Monotone Likelihood Ratio Property

We begin by recalling the definitions of the *monotone likelihood ratio property* (henceforth abbreviated as MLRP) and the *strict monotone likelihood ratio property* (henceforth abbreviated as strict MLRP).

**Definition 1.** *For any totally ordered set $(\mathcal{P}, \preceq)$, $X : \Omega \to \mathcal{P}$ satisfies*

[4]See also "Performance Management: Which Performance Rating Scale is Best, and What Should an Employer Consider in Adopting a Performance Rating Scale?," *The Society of Human Resource Foundation*, Oct 21, 2014.

*(a) The monotone likelihood ratio property with respect to $\preceq$ if for all $A, A' \in \mathcal{P}$, $z(A) \leq z(A')$ if and only if $A \preceq A'$;*

*(b) The strict monotone likelihood ratio property with respect to $\preceq$ if for all $A, A' \in \mathcal{P}$, $z(A) < z(A')$ if and only if $A \preceq A'$.*

**Definition 2.** *Any $A, A' \in \Sigma$ satisfy $A \overset{z}{\preceq} A'$ if and only if $z(A) \leq z(A')$.*

It is worth noting that while the performance measure induced by any arbitrary monitoring technology satisfies the MLRP with respect to $\overset{z}{\preceq}$ by definition,[5] it violates the strict MLRP with respect to $\overset{z}{\preceq}$ in case there are multiple monitoring outcomes that attain the same $z$-value. This observation should be contrasted with the next theorem, which shows that the performance measure induced by the optimal monitoring technology always satisfies the strict MLRP with respect to $\overset{z}{\preceq}$.

**Theorem 1.** *Under Assumption 1, any $\mathcal{P}^*$ can be expressed as $\{A_1, \cdots, A_N\}$ where*

*(i) $z(A_1) < z(A_2) < \cdots < z(A_N)$ and $w^*(A_1) = 0 < w^*(A_2) < \cdots < w^*(A_N)$;*

*(ii) $X : \Omega \to \mathcal{P}^*$ satisfies the strict MLRP with respect to $\overset{z}{\preceq}$.*

Theorem 1 can be shown in two steps. First, we take any arbitrary monitoring technology $\mathcal{P}$ as given and solve for the optimal wage scheme for $\mathcal{P}$, i.e.,

$$\min_{w:\mathcal{P}\to\mathbb{R}_+} \sum_{A\in\mathcal{P}} P_1(A)w(A), \text{ s.t. (IC) and (LL)}.$$

Denote the solution to this problem by $w^*(\cdot;\mathcal{P})$. Define

$$\hat{z} = \frac{1}{u'(0)}.$$

The next lemma restates Holmstrom (1979)'s *sufficient statistics principle*, that $w^*(\cdot;\mathcal{P})$ depends only on the $z$-value of the monitoring outcome.

**Lemma 1.** *For any $\mathcal{P}$, there exists $\lambda > 0$ such that $u'(w^*(A;\mathcal{P})) = 1/\max\{\lambda z(A), \hat{z}\}$ for all $A \in \mathcal{P}$.*

[5] Milgrom (1981) observes that for any totally ordered set $(\mathcal{P}, \preceq)$, $X : \Omega \to \mathcal{P}$ satisfies the MLRP with respect to $\preceq$ if and only if $\preceq$ and $\overset{z}{\preceq}$ are consistent, meaning that for all $A, A' \in \mathcal{P}$, $A \preceq A'$ if and only if $A \overset{z}{\preceq} A'$.

Lemma 1 reiterates the fact that $z$-value contains all the information that the principal needs in order to deter shirking. From this follows that when monitoring is costly and flexible, the principal should focus exclusively on the processing, storage and communication of $z$-values, and ignore the part of performance state that is orthogonal to the $z$-value. In particular, if two monitoring outcomes attain the same $z$-value or yield the same wage, then merging them together has no incentive effect but saves the monitoring cost. Hence all outcomes prescribed by the optimal monitoring technology attain distinct $z$-values, from which strict MLRP follows.

## 3.2 Information Aggregation

We begin with the concept of *Z-convexity*.

**Definition 3.** *A set $A \in \Sigma$ is Z-convex if for all $\omega', \omega'' \in A$,*

$$\{\omega \in \Omega : Z(\omega) = (1-s) \cdot Z(\omega') + s \cdot Z(\omega'') \text{ for some } s \in (0,1)\} \subset A.$$

In words, $Z$-convexity means that if a set contains two states with distinct $z$-values, then it must also contain all states with in-between $z$-values. Notice that when $Z(\Omega)$ is connected in $\mathbb{R}$, the $Z$-convexity of a set $A$ reduces to the convexity of $Z(A)$ in $\mathbb{R}$.

The next theorem shows that all cells of the optimal monitoring technology are $Z$-convex and can be obtained from applying a cutoff rule to the space of $z$-values under mild regularity conditions.

**Theorem 2.** *Under Assumptions 1, each $A \in \mathcal{P}^*$ is Z-convex. If, in addition, that $Z(\Omega)$ is connected, then $\mathcal{P}^*$ can be expressed as $\{A_1, \cdots, A_N\}$, where there exist $-\infty \leq \widehat{z}_0 \leq \widehat{z}_1 \leq \cdots \leq \widehat{z}_N < +\infty$ such that each $A_n$ contains $\{\omega : Z(\omega) \in (\widehat{z}_{n-1}, \widehat{z}_n)\}$ and potentially a subset of $\{\omega : Z(\omega) = \widehat{z}_{n-1} \vee \widehat{z}_n\}$.*

Theorem 2 says that the optimal monitoring technology aggregates the fine-grained and high-dimensional performance states with similar $z$-values into the same coarse and single-dimensional grade. This result explains why multi-source performance appraisal systems compress the various sources of feedback into coarse and rank-ordered ratings, such as "outstanding," "highly effective," "satisfactory" and "unsatisfactory," according to the employee's performance measured by an overall

score. It also justifies the assignment of coarse grades, such as A, B, C, D and F, based on the student's performance in terms of an overall grade.

To better understand this result, consider how we should assign each performance state to the various monitoring outcomes in order to mimimize the sum of compensation cost and monitoring cost. Since performance states with similar $z$-values have similar effects on the (IC) constraint and hence the compensation cost, whereas monitoring is flexible, meaning that the monitoring cost is independent of the likelihood ratios of the monitoring outcomes, it follows that the assignment of performance states should follow an *in-betweenness rule over $z$-values*, whereby if two states with different $z$-values are assigned to the same monitoring outcome, then any state with an in-between $z$-value should be assigned to this monitoring outcome, too. Under mild regularity conditions, this in-betweenness rule can be reduced to the simple cutoff rule described above.

Theorem 2 is another showcase of the sufficient statistics principle. Intuitively, since $z$-value is a sufficient statistic for the agent's performance, the assignment of final performance grade should respect the order over $z$-values when monitoring is costly and yet flexible. It is worth distinguishing this result from the conventional interpretation of the sufficient statistics principle, which says that in the presence of multiple sources of performance signals, such as supervisory reports $\omega_1$ and customer reviews $\omega_2$, it suffices to contract on a subset of signals, say $\omega_1$, if and only if $\omega_1$ is a sufficient statistic for $\omega_2$, or equivalently if $\omega_2$ is redundant given $\omega_1$. This interpretation, if taken seriously, cannot explain why so many resources are spent on the collection and processing of redundant signals if in the end, only one single signal is being used to grade and reward the agent. In contrast, we allow the principal to distill $Z(\omega_1, \omega_2)$ from $(\omega_1, \omega_2)$ and to discard the part of information that is orthogonal to $Z(\omega_1, \omega_2)$. This flexibility gives rise to information aggregation even if $\omega_1$ and $\omega_2$ are non-redundant given each other.

## 3.3 Discussions

**Marginal monitoring cost** The marginal monitoring cost parameterizes the practical challenges that arise in the implementation of fine-grained monitoring technologies. Recent empirical studies by Bloom and Van Reenen (2006, 2007, 2010) identify a list of factors that affect this cost: on the supply side, these authors find that access

to IT, transmission of advanced managerial practices and loose labor market regulation reduce the cost of implementing fine-grained monitoring technologies; on the demand side, they find that tough product market competition increases the demand for high-powered incentive contracts, which in turn reduces the opportunity cost of adopting fine-grained monitoring technologies.

**Optimal degree of fine-grainedness** The fine-grainedness of the optimal monitoring technology hinges on the trade-off between the compensation cost and the monitoring cost. Formally, let

$$W(\mathcal{P}) = \sum_{A \in \mathcal{P}} P_1(A) w^*(A; \mathcal{P})$$

be the minimal compensation cost that is incurred by any given monitoring technology $\mathcal{P}$, and express the optimal incentive contract $\langle \mathcal{P}^*(\mu), w^*(\cdot; \mu) \rangle$ as a function of $\mu$. The next proposition examines how the optimal incentive contract varies with $\mu$.

**Proposition 1.** *For any $0 < \mu' < \mu''$, any $\mathcal{P}^*(\mu')$ and $\mathcal{P}^*(\mu'')$ satisfy $H(\mathcal{P}^*(\mu'), 1) \geq H(\mathcal{P}^*(\mu''), 1)$ and $W(\mathcal{P}^*(\mu')) \leq W(\mathcal{P}^*(\mu''))$.*

Proposition 1 says that as the marginal monitoring cost increases, the principal differentiates the agent's performance less carefully and decreases the power of the wage scheme accordingly. In case $H(\mathcal{P}, a) = f(|\mathcal{P}|)$ for some increasing function $f$, the optimal rating scale is non-increasing in the marginal monitoring cost.

More broadly, Proposition 1 illustrates how the variation in the monitoring cost identified Bloom and Van Reenen (2006, 2007, 2010) can lead to the use of different monitoring technologies among otherwise similar firms. This result adds an explanation to a long-lasting puzzle, that of why heterogeneity prevails among the management practices adopted by firms with similar production technologies (see Gibbons and Henderson (2012) for a survey on this subject matter). In Section 5, we further examine the implication of this heterogeneity for multi-agent models.

**Contingent contract** As in Dye (1986), suppose that after the agent privately exerts an effort $a$, players observe at no cost the realization $s \in S$ of a signal (e.g., the company's cash flow) that is distributed independently of $\omega$ given $a$. Based on $s$, the principal acquires information about $\omega$ through the monitoring technology that she pre-commits to. In this setting, an incentive contract is a profile of contingent

partitions and wage schemes $(\langle \mathcal{P}(s), w(\cdot; s)\rangle)_{s\in S}$, where $\mathcal{P}(s)$ is a finite partition of $\Omega$, and $w(\cdot; s)$ maps each $A \in \mathcal{P}(s)$ to a non-negative wage $w(A; s) \geq 0$ for each $s \in S$. Time evolves as follows:

1. The principal commits to a contingent contract $(\langle \mathcal{P}(s), w(\cdot; s)\rangle)_{s\in S}$;
2. The agent privately exerts an effort $a \in \mathcal{A}$;
3. Nature draws $\omega \in \Omega$ and $s \in S$.
4. Players observe $s$ and the unique cell $A(\omega; s)$ of $\mathcal{P}(s)$ that contains $\omega$;
5. The principal pays the promised wage $w(A(\omega; s); s)$ to the agent.

Since $\omega$ and $s$ are independently distributed given $a$, the introduction of $s$ has no effect on the mapping $Z : \Omega \to \mathbb{R}$ or the $z$-value of $A \in \Sigma$. Let $S$ be finite. Define

$$Z(s) = 1 - \frac{P_0(s)}{P_1(s)}$$

for each $s \in S$, and suppose that $Z(s)$ is non-zero for all $s \in S$.

**Assumption 2.** *$Z(s) \neq 0$ for all $s \in S$.*

The next corollary characterizes the optimal contingent contract.

**Corollary 1.** *Suppose Assumptions 1 and 2 hold. Then for each $s \in S$, any $\mathcal{P}^*(s)$ can be expressed as $\{A_{1,s}, A_{2,s}, \cdots, A_{N(s),s}\}$ where (i) $z(A_{1,s}) < z(A_{2,s}) < \cdots < z(A_{N(s),s})$ and $w^*(A_{1,s}; s) = 0 < w^*(A_{2,s}; s) < \cdots < w^*(A_{N(s),s}; s)$, and (ii) each $A_{n,s}$ is $Z$-convex.*

# 4 Multiple Actions

This section extends the baseline model to encompass multiple deviant actions. Specifically, let $\mathcal{A}$ be any finite set and $a^* \in \mathcal{A}$ be the target action that the principal aims to induce. To make the analysis interesting, suppose that the target action is not the least costly action, i.e., $c(a^*) > \min_{a\in\mathcal{A}} c(a)$. For each deviant action $a \in \mathcal{D} = \mathcal{A} - \{a^*\}$, define a random variable $Z_a : \Omega \to \mathbb{R}$ by

$$Z_a = 1 - \frac{p_a}{p_{a^*}}.$$

For each $a \in \mathcal{D}$ and $A \in \Sigma$, define the $z_a$-value of $A$ by

$$z_a(A) = 1 - \frac{P_a(A)}{P_{a^*}(A)}.$$

A contract is incentive compatible if for each $a \in \mathcal{D}$, we have

$$\sum_{A \in \mathcal{P}} u(w(A)) P_{a^*}(A) z_a(A) \geq c\left(a^*\right) - c(a). \tag{IC$_a$}$$

Take any profile of non-negative reals $\vec{\lambda} = (\lambda_a)_{a \in \mathcal{D}}$. Define a random variable $\overline{Z}\left(\vec{\lambda}\right) : \Omega \to \mathbb{R}$ by

$$\overline{Z}\left(\vec{\lambda}\right) = \sum_{a \in \mathcal{D}} \lambda_a \cdot Z_a.$$

For each $A \in \Sigma$, define the $\overline{z}\left(\vec{\lambda}\right)$-value of $A$ by

$$\overline{z}\left(A; \vec{\lambda}\right) = \sum_{a \in \mathcal{D}} \lambda_a \cdot z_a(A).$$

The next definition generalizes $Z$-convexity.

**Definition 4.** *A set $A \in \Sigma$ is $\overline{Z}\left(\vec{\lambda}\right)$-convex if for all $\omega', \omega'' \in A$,*

$$\left\{\omega : \overline{Z}\left(\omega; \vec{\lambda}\right) = (1-s) \cdot \overline{Z}\left(\omega'; \vec{\lambda}\right) + s \cdot \overline{Z}\left(\omega''; \vec{\lambda}\right) \text{ for some } s \in (0,1)\right\} \subset A.$$

With this definition in hand, we now characterize the optimal incentive contract that deters multiple deviant actions.

**Theorem 3.** *Under Assumption 1, any $\mathcal{P}^*$ can be expressed as $\{A_1, \cdots, A_N\}$, where there exist a profile $\vec{\lambda^*} = (\lambda_a^*)_{a \in \mathcal{D}}$ of non-negative reals with $\max_{a \in \mathcal{D}} \lambda_a^* > 0$ such that*

*(i) $\overline{z}\left(A_1; \vec{\lambda^*}\right) < \overline{z}\left(A_2; \vec{\lambda^*}\right) < \cdots < \overline{z}\left(A_N; \vec{\lambda^*}\right)$ and $w^*\left(A_1; \vec{\lambda^*}\right) = 0 < w^*\left(A_2; \vec{\lambda^*}\right) < \cdots < w^*\left(A_N; \vec{\lambda^*}\right)$;*

*(ii) Each $A_n \in \mathcal{P}^*$ is $\overline{Z}\left(\vec{\lambda^*}\right)$-convex. If, in addition, that $\overline{Z}\left(\Omega; \vec{\lambda^*}\right)$ is connected in $\mathbb{R}$, then there exist $-\infty \leq z_0^* \leq \cdots \leq z_N^* < +\infty$ such that each $A_n$ contains $\left\{\omega : \overline{Z}\left(\omega; \vec{\lambda^*}\right) \in \left(z_{n-1}^*, z_n^*\right)\right\}$ and potentially a subset of $\left\{\omega : \overline{Z}\left(\omega; \vec{\lambda^*}\right) = z_{n-1}^* \vee z_n^*\right\}$.*

According to Theorem 3, the optimal incentive contract that deters multiple deviant actions takes the form of a *balanced scorecard* (Kaplan and Norton (1992, 1993)), whereby agent's performance $Z_a$ in resisting each potential deviation $a \in \mathcal{D}$ is weighed by the Lagrange multiplier $\lambda_a^*$ of the corresponding incentive compatibility constraint. The Lagrange multipliers reveal how the principal trades off the detection of various deviations when monitoring is costly and yet flexible. To illustrate, consider two scenarios where (1) $\lambda_a^*$ is large and (2) $\lambda_a^* = 0$. In the first scenario, the agent is tempted to commit deviation $a$, and the principal best-responds by concentrating on the detection of deviation $a$ and varying the monitoring outcome significantly with the agent's performance $Z_a$ in resisting deviation $a$. In the second scenario, the agent has no desire to commit deviation $a$. Therefore, the principal spends no resource on the detection of deviation $a$ and leaves the monitoring outcome invariant to $Z_a$.

Motivated by the increasing task complexity facilitated by IT, Kaplan and Norton (1992, 1993) invented the balanced scorecard, with the aim of improving employee performance by monitoring and rewarding a range of linked activities. As of today, balanced scorecard is regarded as one of the first types of pay schemes that make use of the greater amount of information due to advances in technology, and variants of it has been adopted by many large firms and organizations across the globe (see Griff and Neely (2009) and the references therein). But since its invention, balanced scorecard has been criticized, most fiercely by Jensen (2001), for giving no clue about how managers should trade off the monitoring of different activities. Theorem 3 closes this debate: this trade-off is determined by how tempted that the employee feels about shirking each activity. In the next section, we examine the implication of this result for agency models with multiple tasks.

## 4.1 Multiple Tasks

As an application, consider an adaptation of Holmstrom and Milgrom (1991)'s multi-task model, where the agent can privately exert either the high effort ($a_i = 1$) or the low effort ($a_i = 0$) in each of the two tasks $i = 1, 2$. Each $a_i$ *independently* generates a probability space $(\Omega_i, \Sigma_i, P_{i,a_i})$, where $\Omega_i \subset \mathbb{R}^{k_i}$ is the state space of the agent's performance in task $i$, $\Sigma_i$ is the Borel sigma-algebra restricted to $\Omega_i$, and $P_{i,a_i}$ is the probability measure over $(\Omega_i, \Sigma_i)$ given $a_i$. Define $\vec{a} = a_1 a_2$ and $\vec{\omega} = \omega_1 \omega_2$. Let $\mathcal{A} = \{11, 01, 10, 00\}$, $a^* = 11$ and $\mathcal{D} = \{01, 10, 00\}$, and define $(Z_{\vec{a}})_{\vec{a} \in \mathcal{D}}$ the same way

as we did in the previous section. Under the assumption that performance states are independently distributed across tasks, we have

$$Z_{01}(\vec{\omega}) = Z_{01}(\omega_1),$$

$$Z_{10}(\vec{\omega}) = Z_{10}(\omega_2),$$

and

$$Z_{00}(\vec{\omega}) = Z_{01}(\omega_1) + Z_{10}(\omega_2) - Z_{01}(\omega_1) \cdot Z_{10}(\omega_2).$$

For any profile $\vec{\lambda} = (\lambda_{\vec{a}})_{\vec{a}\in\mathcal{D}}$ of non-negative reals, define

$$\overline{Z}\left(\vec{\omega}; \vec{\lambda}\right) = (\lambda_{01} + \lambda_{00}) \cdot Z_{01}(\omega_1) + (\lambda_{10} + \lambda_{00}) \cdot Z_{10}(\omega_2) - \lambda_{00} \cdot Z_{01}(\omega_1) \cdot Z_{10}(\omega_2).$$

A straightforward extension of Theorem 3 leads to the following characterization for the optimal multi-task contract.

**Corollary 2.** *Suppose that Assumption 1 holds and that $Z_{\vec{a}}(\Omega_1 \times \Omega_2)$ is connected in $\mathbb{R}$ for all $\vec{a} \in \mathcal{D}$. Then any $\mathcal{P}^*$ can be expressed as $\{A_1, \cdots, A_N\}$, where there exist (a) $\vec{\lambda^*} = (\lambda^*_{\vec{a}})_{\vec{a}\in\mathcal{D}}$ where $\lambda^*_{\vec{a}} \geq 0$, $\lambda^*_{01} + \lambda^*_{00} > 0$ and $\lambda^*_{10} + \lambda^*_{00} > 0$, and (b) $-\infty \leq z^*_0 \leq \cdots \leq z^*_N < +\infty$, such that each $A_n$ contains $\left\{\vec{\omega} : \overline{Z}\left(\vec{\omega}; \vec{\lambda^*}\right) \in \left(z^*_{n-1}, z^*_n\right)\right\}$ and potentially a subset of $\left\{\vec{\omega} : \overline{Z}\left(\vec{\omega}; \vec{\lambda^*}\right) = z^*_{n-1} \vee z^*_n\right\}$.*

Comparing and constrasting Corollary 2 with the result of Holmstrom and Milgrom (1991) yields new insights into the management of multi-task agency relationships. In their seminal paper, Holmstrom and Milgrom (1991) argues that when the agent faces multiple tasks with different measurabilities (i.e., the performance-measuring signals have different noise levels), over-incentivizing the easy-to-measure task prevents the agent from completing the difficult-to-measure task. At the heart of this argument is the following view, that fine-tuning the power of the compensation scheme is key to the management of multi-task agency relationships when the performance-measuring signals can only be taken as exogenously given.

Corollary 2 suggests a conjugate solution, that of fine-tuning the monitoring intensity across tasks according to the agent's tendency to shirk. To illustrate, suppose that the performance state in task one is drawn from a noisy probability distribution. In this case, since the one-step deviation 01 and the double-step deviation 00 are most difficult to detect, it is easy to construct examples where the ratio

$(\lambda_{00}^* + \lambda_{01}^*)/(\lambda_{00}^* + \lambda_{10}^*)$ exceeds one, meaning that it is optimal to focus mostly on the monitoring of task one and to vary the monitoring outcome significantly with the agent's performance in task one. This result has policy implications, including how universities should assess the teaching performance of faculties who simultaneously engage in more difficult-to-measure tasks such as research.

The Lagrange multipliers constitute part of the endogenous solution to the principal's problem. To complete the above story, notice that as the principal shifts focus to the monitoring of task one, the agent becomes more tempted to shirk task two instead, which in turn has two effects on the optimal monitoring technology. The first effect resembles the substitution effect in consumer theory, as requires that the principal re-balance the monitoring intensity across tasks. The second effect is analogous to the income effect in consumer theory, as it stipulates that adjustments be made to the performance thresholds $(z_n^*)_{n=1}^N$. The analytical and numerical results that we have so far obtained (available upon request) suggest that the overall effect depends subtly on model primitives, especially the probability distribution functions of performance states. Thus it is not surprising that Kaplan and Norton (1992, 1993) left the trade-off between different tasks unspecified. This in turn suggests that in practice, a careful evaluation of the contracting environment is required for determining the monitoring intensity for each task.

# 5 Multiple Agents

**Setup** A risk-neutral principal faces multiple risk-averse agents $i = 1, \cdots, I$, each of whom can spend a non-negative wage $w_i \geq 0$ and exert either the high effort ($a_i = 1$) or the low effort ($a_i = 0$), earning a payoff $u_i(w_i) - c_i(a_i)$ that satisfies $u_i(0) = 0$, $u_i' > 0$, $u_i'' < 0$ and $c_i(1) = c_i > c_i(0) = 0$. Each joint effort profile $\vec{a} = (a_1, \cdots, a_I)$ generates a probability space $(\Omega, \Sigma, P_{\vec{a}})$, where $\Omega$ is the space of joint performance states, $\Sigma$ is a sigma-algebra over $\Omega$, and $P_{\vec{a}}$ is the probability measure over $(\Omega, \Sigma)$ given $\vec{a}$. The principal observes neither agent's effort, and her goal is to induce the high effort from all agents.

**Incentive contract** An incentive contract is a pair $\langle \mathcal{P}, \vec{w}(\cdot) \rangle$, where $\mathcal{P}$ is a finite partition of $\Omega$ whose cells belong to $\Sigma$, whereas $\vec{w} : \mathcal{P} \to \mathbb{R}_+^I$ maps each $A \in \mathcal{P}$ to a vector $\vec{w}(A) = (w_1(A), \cdots, w_I(A))$ of non-negative wages. For each joint perfor-

mance state $\vec{\omega}$, let $A(\vec{\omega})$ denote the unique monitoring outcome that contains $\vec{\omega}$, and $\vec{w}(A(\vec{\omega}))$ denote the wage vector at state $\vec{\omega}$. Time evolves as follows:

1. The principal commits to an incentive contract $\langle \mathcal{P}, \vec{w}(\cdot) \rangle$;
2. Agents simultaneously make effort choices $a_i \in \mathcal{A}_i = \{0, 1\}$, $i = 1, \cdots, I$;
3. Nature draws a joint performance state $\vec{\omega}$ from $\Omega$ according to $P_{\vec{a}}$;
4. The monitoring technology publicly announces the monitoring outcome $A(\vec{\omega})$;
5. The principal pays the promised wage $w_i(A(\vec{\omega}))$ to agent $i = 1, \cdots, I$.

Let $\vec{1}$ denote the $I$-dimensional vector of ones. For each $i = 1, \cdots, I$, define a random variable $Z_i : \Omega \to \mathbb{R}$ by

$$Z_i = 1 - \frac{p_{a_i=0, a_{-i}=1}}{p_{\vec{1}}}.$$

For each $i = 1, \cdots, I$ and $A \in \Sigma$, define the $z_i$-value of $A \in \Sigma$ by

$$z_i(A) = \mathbb{E}\left[ Z_i \mid A; \vec{a} = \vec{1} \right].$$

A contract is incentive compatible for agent $i$ if

$$\sum_{A \in \mathcal{P}} u_i(w_i(A)) P_{\vec{1}}(A) z_i(A) \geq c_i, \tag{IC$_i$}$$

and it satisfies agent $i$'s limited liability constraint if

$$w_i(A) \geq 0, \forall A \in \mathcal{P}. \tag{LL$_i$}$$

At any given effort profile $\vec{a}$, the total cost of implementing an incentive contract is given by

$$\sum_{A \in \mathcal{P}} P_{\vec{a}}(A) \sum_{i=1}^{I} w_i(A) + \mu \cdot H(\mathcal{P}, \vec{a}).$$

An optimal incentive contract minimizes the implementation cost, subject to the incentive compatibility constraints and limited liability constraints, i.e.,

$$\min_{\langle \mathcal{P}, \vec{w}(\cdot) \rangle : |\mathcal{P}| \in \mathbb{N}} \sum_{A \in \mathcal{P}} P_{\vec{a}}(A) \sum_{i=1}^{I} w_i(A) + \mu \cdot H(\mathcal{P}, \vec{a}), \text{ s.t. (IC}_i\text{) and (LL}_i\text{)}, i = 1, \cdots, I.$$

**Optimal multi-agent contract** Define a mapping $\vec{Z}:\Omega\to\mathbb{R}^I$ by

$$\vec{Z}=(Z_1,\cdots,Z_I),$$

and use $\vec{Z}(A)$ to denote the image of any $A\in\Sigma$ under $\vec{Z}$. The next definition generalizes $Z$-convexity.

**Definition 5.** *A set $A\in\Sigma$ is $\vec{Z}$-convex if for any $\vec{\omega}',\vec{\omega}''\in A$,*

$$\left\{\vec{\omega}\in\Omega:\vec{Z}(\vec{\omega})=(1-s)\cdot\vec{Z}(\vec{\omega}')+s\cdot\vec{Z}(\vec{\omega}'')\ \textit{for some}\ s\in(0,1)\right\}\subset A.$$

The next assumption is meant for analytical elegance.

**Assumption 3.** *$\vec{Z}(\Omega)$ is connected in $\mathbb{R}^I$ and the distribution of $\vec{Z}$ given $\vec{a}=\vec{1}$ is atomless.*

With these definition and assumption in hand, we now characterize the optimal multi-agent contract.

**Theorem 4.** *Under Assumptions 1, any $\mathcal{P}^*$ satisfies the following properties:*

*(i) For each $i=1,\cdots,I$, there exists $\lambda_i>0$ such that any $A\in\mathcal{P}^*$ satisfies $u_i'(w_i^*(A))=1/\max\{\lambda_i z_i(A),\hat{z}\}$;*

*(ii) Each $A\in\mathcal{P}^*$ is $\vec{Z}$-convex. If, in addition, that Assumption 3 is satisfied, then the boundaries of $A$ consist of straight line segments in $\vec{Z}(\Omega)$.*

## 5.1 Individual vs. Group Incentive Contract

Theorem 4 enables us to compare individual and group incentive contracts from the angle of monitoring cost. In order single out our contribution, we will henceforth work with *technologically independent* agents. Formally, suppose each agent $i$'s individual effort $a_i$ generates a probability space $(\Omega_i,\Sigma_i,P_{i,a_i})$, where $\Omega_i\subset\mathbb{R}^{d_i}$ is the space of agent $i$'s individual performance states, $\Sigma_i$ is the Borel sigma-algebra restricted to $\Omega_i$, and $P_{i,a_i}$ is the probability measure over $(\Omega_i,\Sigma_i)$ given $a_i$. Agents are said to be technologically independent if the probability space generated by the joint effort profile has the following product structure.

**Assumption 4.** $(\Omega, \Sigma, P_{\vec{a}}) = (\Omega_1 \times \cdots \times \Omega_I, \Sigma_1 \otimes \cdots \otimes \Sigma_I, P_{1,a_1} \times \cdots \times P_{I,a_I})$ *for all* $\vec{a} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_I$.

In the language of contract theory, Assumption 4 rules out any kind of *technological linkage* (i.e., $\omega_i$ depends on $a_{-i}$) or *common productivity shock* (i.e., $\omega_i, \omega_j$ are correlated given $\vec{a}$) between agents (see Segal (2006) for terminologies).

### 5.1.1 Partitional Representation

The analysis below considers both finite and infinite partitions. We begin with a few definitions.

**Definition 6.** *A partition* $\mathcal{P}$ *of* $\Omega$ *is an* individual monitoring technology *if for every* $A \in \mathcal{P}$*, there exist* $A_i \in \Sigma_i$ *such that* $A = A_1 \times \cdots \times A_I$*; otherwise it is a* group monitoring technology.

**Definition 7.** *Under any individual monitoring technology* $\mathcal{P}$*, a wage scheme* $\vec{w}(\cdot; \mathcal{P})$ *for* $\mathcal{P}$ *is an* individual wage scheme *if* $w_i(A_i \times A'_{-i}; \mathcal{P}) = w_i(A_i \times A''_{-i}; \mathcal{P})$ *for all* $i = 1, \cdots, I$ *and* $A_i \times A'_{-i}, A_i \times A''_{-i} \in \mathcal{P}$.

**Definition 8.** $\langle \mathcal{P}, \vec{w}(\cdot; \mathcal{P}) \rangle$ *is an* individual incentive contract *if* $\mathcal{P}$ *is a individual monitoring technology and* $\vec{w}(\cdot; \mathcal{P})$ *is an individual wage scheme for* $\mathcal{P}$*; otherwise it is a* group incentive contract.

The next lemma shows that coupling the use of individual monitoring technology with that of individual wage scheme is optimal when agents are technologically independent.

**Lemma 2.** *Under Assumption 4,* $\langle \mathcal{P}, \vec{w}^*(\cdot; \mathcal{P}) \rangle$ *is an* individual incentive contract *for any individual monitoring technology* $\mathcal{P}$.

Theorem 4 and Lemma 2 enable us to represent individual and group incentive contracts as partitions of the space of $\vec{Z}(\Omega)$. Figures 2 and 3 give such examples in case $I = 2$. In particular, the individual incentive contract depicted in Figure 2 has the minimal cardinality that any individual incentive contract needs in order to elicit the high effort from both agents. Meanwhile, Figure 3 exhausts all bi-partitional contracts, where the *team* depicted on the left panel evaluates and rewards both agents together, whereas the *tournament* depicted on the right panel conducts

relative performance evaluations and rewards the best performer only. In the current framework, all these contracts can be adopted at potentially different monitoring costs.

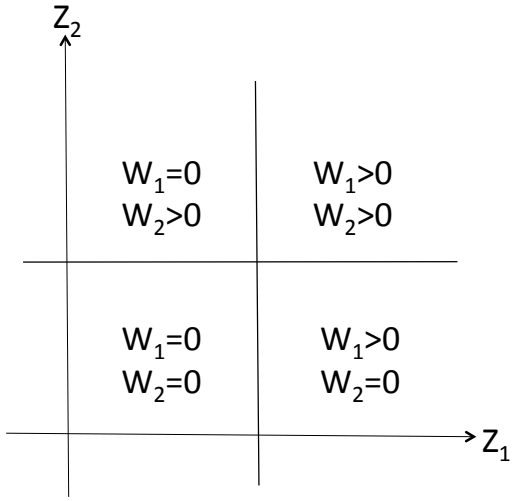


Figure 2: Individual incentive contract.

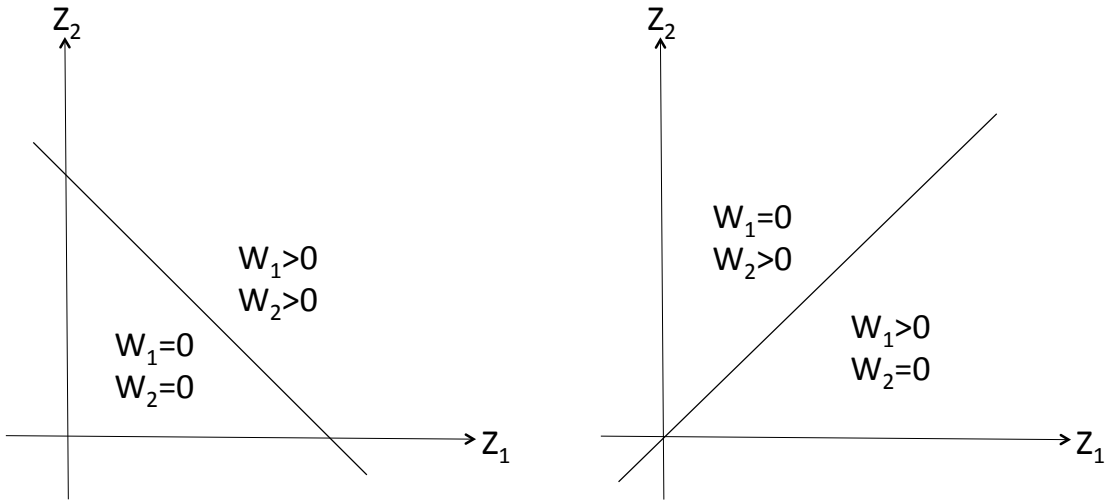


Figure 3: Team and tournament.

Since Lazear and Rosen (1981), it has long been noted that individual performance measures contain more fine-grained information than ordinal performance rankings.[6]

[6]A quote from Lazear and Rosen (1981) goes: "*In a modern, complex business organization ... the costs of measurement for each conceivable candidate are prohibitively expensive. Instead, it might*

In the survey conducted by Bloom and Van Reenen (2006, 2007), rewarding employees based on their individual performance rather than the shift performance or the overall company performance is regarded as a good management practice. A close inspection of Figures 2 and 3 leads to a similar conclusion, that group monitoring technologies lump the assessment of agents together and generate coarser performance ratings than individual monitoring technologies. The next lemma formalizes this intuition in case the fine-grainedness measure is given by the number of performance categories.

**Lemma 3.** *Under Assumptions 3 and 4 hold, for any individual incentive contract that elicits the high effort from both agents, there exists a group incentive contract that does the same through a monitoring technology with a smaller cardinality.*

### 5.1.2 Coexistence of individual and group contract

In their seminal work that ignores the monitoring cost, Holmstrom (1982), Green and Stokey (1983) and Mookherjee (1984) argue that it is optimal to use individual incentive contracts among technologically independent agents. The reason is straightforward: in this case, since the individual performance of an agent is a sufficient statistic for his individual effort, it follows from the sufficient statistics principle that the optimal individual incentive contract minimizes agents' exposure to risks and hence the compensation cost.

Since the discovery of this result, a large body of the agency literature has adopted a dichotomous view that attributes the use of group incentive contracts to either the technological linkage or the common productivity shock between agents on the one hand, and confines the use of individual incentive contracts among technologically independent agents on the other hand. Studies in this vein include Hamilton et al. (2003) and Boning et al. (2007), which quantify the value of team incentive when the task requires diverse skills, frequent communication and collaborative work; as well as Holmstrom (1982), Green and Stokey (1983), Nalebuff and Stiglitz (1983) and Mookherjee (1984), which establishes rank-order tournament as the optimal incentive scheme in case agents face common productivity shocks.

Recently, this conventional wisdom has been challenged by Bloom and Van Reenen (2006, 2007), who find that even firms with similar production technologies make

*be said that those in the running are "tested" by assessments of performance at lower positions. Such tests are inherently ordinal in nature ... It is in these situations such as this that the conditions seem ripe for tournaments to be the dominant incentive contract institution.*"

signfiicantly different choices between individual and group incentive contracts. The next proposition resolves this puzzle from the angle of monitoring cost.

**Proposition 2.** *Suppose that Assumptions 3 and 4 hold, and that* $H(\mathcal{P}, a) = f(|\mathcal{P}|)$ *for some increasing function* $f : \mathbb{N} \cup \{+\infty\} \to \mathbb{R}$.

*(i) When* $\mu = 0$*, the optimal individual incentive contract is an optimal contract;*

*(ii) There exists* $\underline{\mu} > 0$ *such that for all* $\mu > \underline{\mu}$*, any optimal contract is a group incentive contract.*

Proposition 2 illustrates how the variation in the marginal monitoring cost can explain the coexistence of individual and group incentive contracts among technologically independent agents. On the one hand, Part (i) of this proposition replicates the result of Holmstrom (1979), saying that the optimal individual incentive contract minimizes the compensation cost among technologically independent agents. On the other hand, Part (ii) shows that when monitoring is costly and yet flexible, meaning that any partitional monitoring technology can be adopted at a cost measured by its cardinality, the limited monitoring capacity creates an *attentional linkage* between agents that favors group incentive contracts over individual incentive contracts. When the marginal monitoring cost is sufficiently high, saving the monitoring cost becomes the primary concern and gives rise to the use of group incentive contracts among technologically independent agents. The main prediction of this result, that agents are less recognized for their individual performance as the marginal monitoring cost increases, other things equal, is supported by the findings of Bloom and Van Reenen (2006, 2007).

# 6 Conclusion

We conclude by discussing related issues and suggesting avenues for future research. So far, we have allowed the principal to choose between a large variety of monitoring technologies and have made only a few assumptions about the monitoring cost funciton, in order to capture the flexibility in employee monitoring and to obtain robust predictions about employee productivity and the internal organization of firms. When applying our framework to the study of more concrete problems, richer comparative statics can be obtained from imposing more restrictions on either the principal's choice set or the monitoring cost function that best reflect the constraints in reality.

The foregoing analysis assumes that the agent is protected by limited liability and pays no monitoring cost. In Appendices B.2 and B.3, we replace the limited liability constraint with an individual rationality constraint and let the agent pay part of the monitoring cost (e.g., communication cost).

So far, we have ruled out random monitoring technologies for two reasons. First, randomization creates confusion and disputes in human resource management. Second, most of our results, in particular the in-betweenness property of the information aggregation rule, remain valid even if randomization is allowed. See Appendix B.4 for further discussions and results.

Recent advancement in incentive theory (see, for example, Li (2015)) has attempted to identify which features of the monitoring technology have the most significant impact for the efficiency of long-term agency relationships. In light of these results, it is imperative that we develop a framework for studying the optimal choice of monitoring technology in dynamic agency models with moral hazard. We hope the current analysis provides a useful starting point for such analysis.

# A Omitted Proofs

For brevity, write $\pi_n$, $z_n$, $w_n$ and $w_n^*$ instead of $P_1(A_n)$, $z(A_n)$, $w(A_n)$ and $w^*(A_n)$.

Proof of Lemma 1.

*Proof.* Take $\mathcal{P}^* = \{A_1, \cdots, A_N\}$ as given and reduce the principal's problem to

$$\begin{aligned} &\min_{\{w_n\}} \sum_{n=1}^{N} \pi_n w_n, \\ &\text{s.t. } \sum_{n=1}^{N} \pi_n u(w_n) z_n \geq c, && \text{(IC)} \\ &\text{and } w_n \geq 0, n = 1, \cdots, N. && \text{(LL)} \end{aligned}$$

Let $\lambda$ and $\eta_n$ denote the Lagrange multiplier associated with the (IC) constraint and the (LL) constraint with respect to $w_n$, respectively. In the principal's problem, taking derivative with respect to $w_n$ yields

$$u'(w_n^*) = \frac{1 - \eta_n/\pi_n}{\lambda z_n}.$$

From this follows that $w_n^* > 0 \iff u'(w_n^*) = 1/\lambda z_n \iff \lambda z_n > \hat{z}$, or equivalently that $u'(w_n^*) = 1/\max\{\lambda z_n, \hat{z}\}$. □

Proof of Theorem 1.

*Proof.* Suppose, to the contrary, that $w_j^* = w_k^*$ for some $j \neq k$. Then by merging $A_j$ and $A_k$ into a single cell, the principal can save the monitoring cost without affecting the (IC) constraint, the (LL) constraint or the compensation cost, a contradiction. Thus $w_j^* \neq w_k^*$ for all $j \neq k$. Furthermore, there exists $z_n < 0 < \hat{z}$, because $\sum_{n=1}^N \pi_n z_n = 0$. Together, these results enable us to express $\mathcal{P}^*$ as $\{A_1, \cdots, A_N\}$ where $z_1 < z_2 < \cdots < z_N$ and $w_1^* = 0 < w_2^* < \cdots < w_N^*$. □

Proof of Theorem 2.

*Proof.* Suppose, to the contrary, that some $A_j \in \mathcal{P}^*$ is not $Z$-convex. Then there exist $A', A'', \tilde{A} \in \Sigma$ such that

(1) $P_1(A') = P_1(A'') = P_1\left(\tilde{A}\right) = \epsilon$ for some small $\epsilon > 0$;

(2) $A', A'' \subset A_j$ and $\tilde{A} \subset A_k \in \mathcal{P}^*$ for some $k \neq j$;

(3) $z\left(\tilde{A}\right) = (1-s)z(A') + sz(A'')$ for some $s \in (0,1)$.

For brevity, write $\tilde{z} = z\left(\tilde{A}\right)$, $z' = z(A')$ and $z'' = z(A'')$.

Consider two perturbations to $\mathcal{P}^*$:

(a) Moving $A'$ to $A_k$ and $\tilde{A}$ to $A_j$ (henceforth referred to as "*switching*" $A'$ and $\tilde{A}$);

(b) Moving $\tilde{A}$ to $A_j$ and $A''$ to $A_k$ (henceforth referred to as "*switching*" $\tilde{A}$ and $A''$).

Under Assumption 1, neither perturbation affects the monitoring cost. We now argue that one of these perturbations reduces the compensation cost without violating the (IC) or the (LL) constraint.

Consider perturbation (a) first. Specifically, let $(z_n(\epsilon))_{n=1}^N$ denote the $z$-values of the cells of $\mathcal{P}^*$ after this perturbation. Straightforward algebra shows that

$$\begin{cases} z_j'(0) = \dfrac{s(z''-z')}{\pi_j}, \\ z_k'(0) = -\dfrac{s(z''-z')}{\pi_k}, \\ z_n'(0) = 0, \forall n \neq j,k. \end{cases}$$

Take any profile $(w_n(\epsilon))_{n=1}^N$ of wages (which clearly exists) that satisfies two conditions: (1) $w_1(\epsilon) = w_1(0) = 0$, and (2) (IC) holds after perturbation (a), i.e.,

$$\sum_{n=1}^{N} \pi_n u(w_n(\epsilon)) z_n(\epsilon) = c.$$

In this new (IC) constraint, taking total derivative with respect to $\epsilon$ and multiplying the result by $\lambda$ (the Lagrange multiplier associated with the (IC) constraint prior to the perturbation) yields

$$\begin{aligned}\sum_{n=1}^{N} \pi_n \cdot u'\left(w_n^*\right) \cdot \lambda z_n \cdot w_n'(0) &= -\lambda \left[u\left(w_j^*\right) \cdot \pi_j z_j'(0) + u\left(w_k^*\right) \cdot \pi_k z_k'(0)\right] \\ &= s\left[u\left(w_k^*\right) - u\left(w_j^*\right)\right]\left(\lambda z'' - \lambda z'\right).\end{aligned}$$

Now since $u'\left(w_n^*\right) = \frac{1}{\lambda z_n}$ for all $n \geq 2$ and $w_1'(0) = 0$, it follows that

$$u'\left(w_n^*\right) \cdot \lambda z_n \cdot w_n'(0) = w_n'(0), \forall n = 1, \cdots, N.$$

Plugging in this condition into the previous one yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = s\left[u\left(w_k^*\right) - u\left(w_j^*\right)\right]\left(\lambda z'' - \lambda z'\right), \tag{A.1}$$

where the left-hand side equals the rate of change in the compensation cost.

Now consider perturbation (b). Similar algebraic manipulation yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = -(1-s)\left[u\left(w_k^*\right) - u\left(w_j^*\right)\right]\left(\lambda z'' - \lambda z'\right). \tag{A.2}$$

Since $u\left(w_j^*\right) \neq u\left(w_k^*\right)$, the right-hand sides of (A.1) and (A.2) have the opposite signs. Thus for one of perturbations (a) and (b), we can construct a wage profile such that after the perturbation, the compensation cost decreases while both (IC) and (LL) remain satisfied. Thus $\mathcal{P}^*$ is not optimal, a contradiction. □

Proof of Proposition 1.

*Proof.* Take any $0 < \mu' < \mu''$. Since $\mathcal{P}^*\left(\mu'\right)$ and $\mathcal{P}^*\left(\mu''\right)$ are optimal at $\mu = \mu'$ and $\mu =$

$\mu''$, it follows that $W\left(\mathcal{P}^*\left(\mu'\right)\right)+\mu'\cdot H\left(\mathcal{P}^*\left(\mu'\right),1\right)\leq W\left(\mathcal{P}^*\left(\mu''\right)\right)+\mu'\cdot H\left(\mathcal{P}^*\left(\mu''\right),1\right)$, and that $W\left(\mathcal{P}^*\left(\mu''\right)\right)+\mu''\cdot H\left(\mathcal{P}^*\left(\mu''\right),1\right)\leq W\left(\mathcal{P}^*\left(\mu'\right)\right)+\mu''\cdot H\left(\mathcal{P}^*\left(\mu'\right),1\right)$. Telescoping yields the result. □

Proof of Corollary 1.

*Proof.* For each $s\in S$ and $A_{n,s}\in\mathcal{P}^*\left(s\right)$ (the $n^{th}$ cell of partition $s$), let $z_{n,s}$, $w^*_{n,s}$ and $\pi_{n,s}$ denote the $z$-value of $A_{n,s}$, the optimal wage at $A_{n,s}$ and the probability measure of $A_{n,s}$ given $a=1$, respectively.

Part (i): take $\left(\mathcal{P}^*(s)\right)_{s\in S}$ as given and reduce the principal's problem to the following:

$$\begin{aligned}&\min_{\{w_{n,s}\}}\sum_{n\in\mathbb{N},s\in S}P_1(s)\pi_{n,s}w_{n,s},\\&\text{s.t.}\sum_{n\in\mathbb{N},s\in S}P_1(s)\pi_{n,s}u(w_{n,s})z_{n,s}Z(s)\geq c,&\text{(IC)}\\&\text{and } w_{n,s}\geq 0,\forall n\in\mathbb{N},s\in S.&\text{(LL)}\end{aligned}$$

Let $\lambda$ and $\eta_{n,s}$ denote the Lagrange multiplier associated with the (IC) constraint and the (LL) constraint with respect to $w_{n,s}$, respectively. In the principal's problem, taking derivative with respect to $w_{n,s}$ yields

$$u'\left(w^*_{n,s}\right)=\frac{1}{\max\left\{\lambda z_{n,s}Z(s),\hat{z}\right\}},\forall n,s.$$

Plugging this first-order condition into the proof of Theorem 1 yields the result.

Part (ii): suppose, to the contrary, that $A_{j,s}\in\mathcal{P}^*(s)$ is not $Z$-convex for some $j\in\mathbb{N}$ and $s\in S$. Then there exist $A',A'',\tilde{A}\in\Sigma$ such that

(1) $P_1\left(A'\right)=P_1\left(A''\right)=P_1\left(\tilde{A}\right)=\epsilon$ for some small $\epsilon>0$;

(2) $A',A''\subset A_{j,s}$ and $\tilde{A}\subset A_{k,s}\in\mathcal{P}^*\left(s\right)$ for some $k\neq j$;

(3) $z\left(\tilde{A}\right)=(1-t)z\left(A'\right)+tz\left(A''\right)$ for some $t\in(0,1)$.

For brevity, write $z\left(A'\right)=z'$, $z\left(A''\right)=z''$ and $z\left(\tilde{A}\right)=\tilde{z}$.

First, consider the perturbation that switches $A'$ and $\tilde{A}$ at $s$. Let $\left(z_{n,s'}\left(\epsilon\right)\right)_{n\in\mathbb{N},s'\in S}$ denote the $z$-values of the cells of $\mathcal{P}^*\left(s'\right), s'\in S$ after this perturbation. Straightforward algebra shows that

$$\begin{cases} z'_{j,s}\left(0\right)=\dfrac{t\left(z''-z'\right)}{\pi_{j,s}}, \\ z'_{k,s}\left(0\right)=-\dfrac{t\left(z''-z'\right)}{\pi_{k,s}}, \\ z'_{n,s}\left(0\right)=0, \forall n\neq j,k, \\ z'_{n,s'}\left(0\right)=0, \forall n \text{ and } s'\neq s. \end{cases}$$

Take any profile $\left(w_{n,s'}\left(\epsilon\right)\right)_{n\in\mathbb{N},s'\in S}$ of wages (which clearly exists) that satisfies two conditions: (1) $w_{1,s'}(\epsilon)=w_{1,s'}(0)=0$ for all $s'\in S$, and (2) (IC) holds after the perturbation, i.e.,

$$\sum_{s',n}\pi_{n,s'}P_1\left(s'\right)u\left(w_{n,s'}\left(\epsilon\right)\right)z_{n,s'}\left(\epsilon\right)Z\left(s'\right)=c.$$

In this new (IC) constraint, taking total derivative with respect to $\epsilon$ and multiplying the result by $\lambda$ yields

$$\begin{aligned} &\sum_{s',n}\pi_{n,s'}P_1\left(s'\right)\cdot u'\left(w^*_{n,s'}\right)\cdot\lambda z_{n,s'}Z\left(s'\right)\cdot w'_{n,s'}\left(0\right) \\ =&-\lambda P_1\left(s\right)Z\left(s\right)\left[u\left(w^*_{j,s}\right)\cdot\pi_{j,s}z'_{j,s}\left(0\right)+u\left(w^*_{k,s}\right)\cdot\pi_{k,s}z'_{k,s}(0)\right] \\ =&tP_1\left(s\right)Z\left(s\right)\left[u\left(w^*_{k,s}\right)-u\left(w^*_{j,s}\right)\right]\left(\lambda z''-\lambda z'\right). \end{aligned}$$

Since $u'\left(w^*_{n,s'}\right)=\frac{1}{\lambda z_{n,s'}Z(s')}$ for all $n\geq 2$ and $w'_{1,s'}\left(0\right)=0$ for all $s'\in S$, it follows that

$$u'\left(w^*_{n,s'}\right)\cdot\lambda z_{n,s'}Z\left(s'\right)\cdot w'_{n,s'}\left(0\right)=w'_{n,s'}\left(0\right), \forall n\in\mathbb{N}, s'\in S.$$

Substituting this into the previous condition yields

$$\begin{aligned} &\sum_{s',n}\pi_{n,s'}P_1\left(s'\right)w'_{n,s'}\left(0\right) \\ =&tP_1\left(s\right)Z\left(s\right)\left[u\left(w^*_{k,s}\right)-u\left(w^*_{j,s}\right)\right]\left(\lambda z''-\lambda z'\right), \end{aligned} \tag{A.3}$$

where the left-hand side equals the rate of change in the compensation cost.

Second, consider the perturbation that switches $A''$ and $\tilde{A}$ at $s$. Similar algebraic manipulation yields

$$\begin{aligned} &\sum_{s',n} \pi_{n,s'} P_1(s') w'_{n,s'}(0) \qquad (A.4) \\ =& -(1-t) P_1(s) Z(s) \left[u\left(w^*_{k,s}\right) - u\left(w^*_{j,s}\right)\right] (\lambda z'' - \lambda z'). \end{aligned}$$

Since $w^*_{j,s} \neq w^*_{k,s}$ and $Z(s) \neq 0$ (Assumption 2), the right-hand sides of (A.3) and (A.4) have the opposite signs. Thus for one of the above described perturbations, we can construct a wage profile such that after this perturbation, the compensation cost decreases while (IC) and (LL) remain satisfied. Thus $(\mathcal{P}^*(s'))_{s' \in S}$ is not optimal, a contradiction. □

Proof of Theorem 3.

*Proof.* Part (i): take $\mathcal{P}^* = \{A_1, \cdots, A_N\}$ as given and reduce the principal's problem to the following:

$$\begin{aligned} &\min_{\{w_n\}} \sum_{n=1}^{N} \pi_n w_n, \\ &\text{s.t.} \sum_{n=1}^{N} \pi_n u(w_n) z_{a,n} \geq c(a^*) - c(a), \forall a \in \mathcal{D}, \qquad (\text{IC}_a) \\ &\text{and } w_n \geq 0, \forall n = 1, \cdots, N. \qquad (\text{LL}) \end{aligned}$$

Let $\vec{\lambda^*} = (\lambda^*_a)_{a \in \mathcal{D}}$ denote the Lagrange multipliers for the incentive compatibility constraints. Define

$$\mathcal{B} = \{a \in \mathcal{D} : \lambda^*_a > 0\}$$

as the set of deviant actions that attains a binding ($\text{IC}_a$) constraint, and notice that $\mathcal{B} \neq \emptyset$. In the principal's problem, taking derivative with respect to $w_n$ yields

$$u'(w^*_n) = \frac{1}{\max\left\{\overline{z}_n\left(\vec{\lambda^*}\right), \hat{z}\right\}},$$

where $\overline{z}_n\left(\vec{\lambda^*}\right) = \sum_{a \in \mathcal{D}} \lambda^*_a \cdot z_{a,n}$. A straightforward extension of Theorem 2 shows that $\mathcal{P}^*$ can be expressed as $\{A_1, \cdots, A_N\}$ where $\overline{z}_1\left(\vec{\lambda^*}\right) < \cdots < \overline{z}_N\left(\vec{\lambda^*}\right)$ and

$w_1^* = 0 < w_2^* < \cdots < w_N^*$.

Part (ii): suppose, to the contrary, that some $A_j \in \mathcal{P}^*$ is not $\overline{Z}\left(\vec{\lambda^*}\right)$-convex. Then there exist $A', A'', \tilde{A} \in \Sigma$ satisfying

(1) $P_1(A') = P_1(A'') = P_1\left(\tilde{A}\right) = \epsilon$ for some small $\epsilon > 0$;

(2) $A', A'' \subset A_j$ whereas $\tilde{A} \subset A_k \in \mathcal{P}^*$ for some $k \neq j$;

(3) $\overline{z}\left(\tilde{A}; \vec{\lambda^*}\right) = (1-s)\overline{z}\left(A'; \vec{\lambda^*}\right) + s\overline{z}\left(A''; \vec{\lambda^*}\right)$ for some $s \in (0,1)$.

For brevity, write $\overline{z}'\left(\vec{\lambda^*}\right) = \overline{z}\left(A'; \vec{\lambda^*}\right)$, $\overline{z}'\left(\vec{\lambda^*}\right) = \overline{z}\left(A''; \vec{\lambda^*}\right)$ and $\tilde{\overline{z}}\left(\vec{\lambda^*}\right) = \overline{z}\left(\tilde{A}; \vec{\lambda^*}\right)$.

First, consider the perturbation that switches $A'$ and $\tilde{A}$. Let $(z_{a,n}(\epsilon))_{n\in\mathbb{N}, a\in\mathcal{D}}$ denote the $z$-values of the cells of $\mathcal{P}^*$ after this perturbation. Straightforward algebra shows that for each $a \in \mathcal{D}$, we have

$$\begin{cases} z_{a,j}(\epsilon) = \dfrac{s\left(z_a'' - z_a'\right)}{\pi_j} \cdot \epsilon + \mathcal{O}(\epsilon^2), \\ z_{a,k}(\epsilon) = -\dfrac{s\left(z_a'' - z_a'\right)}{\pi_k} \cdot \epsilon + \mathcal{O}(\epsilon^2), \\ z_{a,n}(\epsilon) = z_{a,n}, \forall n \neq j, k. \end{cases}$$

Take any profile $(w_n(\epsilon))_{n=1}^N$ of wages (which clearly exists) that satisfies three conditions: (1) $w_1(\epsilon) = w_1(0) = 0$, (2) any $(\mathrm{IC}_a)$ that used to be slack before the perturbation remains slack after the perturbation, i.e.,

$$\sum_{n=1}^{N} \pi_n u(w_n(\epsilon)) z_{a,n}(\epsilon) > c\left(a^*\right) - c(a), \forall a \in \mathcal{B}^c,$$

and (3) any $(\mathrm{IC}_a)$ that used to be binding before the perturbation becomes weakly slack after the perturbation, i.e.,

$$\sum_{n=1}^{N} \pi_n u(w_n(\epsilon)) z_{a,n}(\epsilon) \geq c\left(a^*\right) - c(a), \forall a \in \mathcal{B}.$$

Multiplying these inequalities by their respective Lagrange multipliers and summing

up the results yields

$$\sum_{n=1}^{N} \pi_n u(w_n(\epsilon)) \overline{z}\left(\epsilon; \vec{\lambda^*}\right) \geq \sum_{a \in \mathcal{D}} \lambda_a \cdot \left(c\left(a^*\right) - c(a)\right).$$

Taking total derivative with respect to $\epsilon$ yields

$$\epsilon \cdot \left[\sum_{n=1}^{N} \pi_n \cdot u'\left(w_n^*\right) \cdot \overline{z}_n\left(\vec{\lambda^*}\right) \cdot w_n'(0) + s\left(u\left(w_j^*\right) - u\left(w_k^*\right)\right)\left(\overline{z}''\left(\vec{\lambda^*}\right) - \overline{z}'\left(\vec{\lambda^*}\right)\right)\right] + \mathcal{O}\left(\epsilon^2\right) \geq 0.$$

Since $u'\left(w_n^*\right) = \frac{1}{\overline{z}_n\left(\vec{\lambda^*}\right)}$ for all $n \geq 2$ and $w_1'(0) = 0$, it follows that

$$u'\left(w_n^*\right) \cdot \overline{z}_n\left(\vec{\lambda^*}\right) \cdot w_n'(0) = w_n'(0), \forall n \in \mathbb{N}.$$

Plugging this into the previous inequality yields

$$\epsilon \cdot \sum_{n=1}^{N} \pi_n w_n'(0) \geq s\left(u\left(w_k^*\right) - u\left(w_j^*\right)\right)\left(\overline{z}''\left(\vec{\lambda^*}\right) - \overline{z}'\left(\vec{\lambda^*}\right)\right) \cdot \epsilon + \mathcal{O}\left(\epsilon^2\right). \tag{A.5}$$

Second, consider the perturbation that switches $A''$ and $\tilde{A}$. Similar algebraic manipulation shows that

$$\epsilon \cdot \sum_{n=1}^{N} \pi_n w_n'(0) \geq -(1-s)\left(u\left(w_k^*\right) - u\left(w_j^*\right)\right)\left(\overline{z}''\left(\vec{\lambda^*}\right) - \overline{z}'\left(\vec{\lambda^*}\right)\right) \cdot \epsilon + \mathcal{O}\left(\epsilon^2\right). \tag{A.6}$$

Since $s \cdot \left(u\left(w_k^*\right) - u\left(w_j^*\right)\right) \cdot \left(\overline{z}''\left(\vec{\lambda^*}\right) - \overline{z}'\left(\vec{\lambda^*}\right)\right)$ and $-(1-s) \cdot \left(u\left(w_k^*\right) - u\left(w_j^*\right)\right) \cdot \left(\overline{z}''\left(\vec{\lambda^*}\right) - \overline{z}'\left(\vec{\lambda^*}\right)\right)$ have the opposite signs, it follows that for one of the above described perturbations, we can construct a wage profile such that after this perturbation, the expected wage decreases whereas all ($\text{IC}_a$) and (LL) remain satisfied. Thus, the contract prior to the perturbation is not optimal, a contradiction. □

Proof of Corollary 2.

*Proof.* $\lambda_{01}^* + \lambda_{00}^* > 0$ and $\lambda_{10}^* + \lambda_{00}^* > 0$ because the contract induces the high effort in both tasks. The rest of the proof follows that of Theorem 3. □

Proof of Theorem 4.

*Proof.* Part (i): write $z_{i,n}$, $w_{i,n}$ and $w^*_{i,n}$ instead of $z_i(A_n)$, $w_i(A_n)$ and $w^*_i(A_n)$ for brevity. Take $\mathcal{P}^* = \{A_1, \cdots, A_N\}$ as given and reduce the principal's problem to

$$\begin{aligned} &\min_{\{w_{i,n}\}} \sum_{n=1}^{N} \pi_n \sum_{i=1}^{I} w_{i,n}, \\ &\text{s.t. } \sum_{n=1}^{N} \pi_n u_i(w_{i,n}) z_{i,n} \geq c_i, \forall i, && (\text{IC}_i) \\ &\text{and } w_{i,n} \geq 0, \forall i, n. && (\text{LL}_i) \end{aligned}$$

Let $\lambda_i$ and $\eta_{i,n}$ denote the Lagrange multiplier for the $(\text{IC}_i)$ constraint and the $(\text{LL}_i)$ constraint with respect to $w_{i,n}$, respectively. In the principal's probelm, taking derivative with respect to $w_{i,n}$ yields

$$u_i'\left(w^*_{i,n}\right) = \frac{1}{\max\left\{\lambda_i z_{i,n}, \hat{z}\right\}}.$$

For each $i = 1, \cdots, I$, let $\mathcal{B}_i$ denote the cells of $\mathcal{P}^*$ that yield a positive wage to agent $i$, i.e.,

$$\mathcal{B}_i = \left\{n : w^*_{i,n} > 0\right\}.$$

Part (ii): suppose, to the contrary, that some $A_j \in \mathcal{P}^*$ is not $\vec{Z}$-convex. Then there exist $A', A'', \tilde{A} \in \Sigma$ such that

(1) $P_1(A') = P_1(A'') = P_1(\tilde{A}) = \epsilon$ for some small $\epsilon > 0$;

(2) $A', A'' \subset A_j$ whereas $\tilde{A} \subset A_k$ for some $k \neq j$;

(3) $\vec{z}\left(\tilde{A}\right) = (1-s) \cdot \vec{z}(A') + s \cdot \vec{z}(A'')$ for some $s \in (0,1)$.

For brevity, write $\vec{z'} = \vec{z}(A')$, $\vec{z''} = \vec{z}(A'')$ and $\vec{\tilde{z}} = \vec{z}\left(\tilde{A}\right)$.

Consider first the perturbation that switches $A'$ and $\tilde{A}$. Let $\left(\vec{z}_n(\epsilon)\right)_{n=1}^N$ denote the

$\vec{z}$-values of the cells of $\mathcal{P}^*$ after this perturbation, where

$$\begin{cases} \vec{z}'_j(0) = \dfrac{s \cdot \left(\vec{z''} - \vec{z'}\right)}{\pi_j}, \\ \vec{z}'_k(0) = -\dfrac{s \cdot \left(\vec{z''} - \vec{z'}\right)}{\pi_k}, \\ \vec{z}'_n(0) = \vec{0}, \forall n \neq j, k. \end{cases}$$

Take any profile $(\vec{w}_n(\epsilon))_{n=1}^N$ of wages (which clearly exists) such that (1) $w_{i,n}(\epsilon) = 0$ for all $i$ and $n \in \mathcal{B}_i^c$, and (2) ($\text{IC}_i$) holds for all $i$ after the perturbation, i.e.,

$$\sum_{n=1}^{N} \pi_n u_i(w_{i,n}(\epsilon)) z_{i,n}(\epsilon) = c_i, \forall i = 1, \cdots, I.$$

In this new ($\text{IC}_i$) constraint, taking total derivative with respect to $\epsilon$ and multiplying the result by $\lambda_i$ yields

$$\begin{aligned} &\sum_{n=1}^{N} \pi_n \cdot u_i'\left(w_{i,n}^*\right) \cdot \lambda_i z_{i,j} \cdot w_{i,n}'(0) \\ =& -\lambda_i \left[u_i\left(w_{i,j}^*\right) \cdot \pi_j z_{i,j}'(0) + u_i\left(w_{i,k}^*\right) \cdot \pi_k z_{i,k}'(0)\right], \forall i = 1, \cdots, I. \end{aligned}$$

Since $u_i'\left(w_{i,n}^*\right) = \frac{1}{\lambda_i z_{i,n}}$ for all $n \in \mathcal{B}_i$ and $w_{i,n}'(0) = 0$ for all $n \in \mathcal{B}_i^c$, it follows that

$$u_i'\left(w_{i,n}^*\right) \cdot \lambda_i z_{i,n}(0) \cdot w_{i,n}'(0) = w_{i,n}'(0), \forall n, i.$$

Plugging this into the previous condition and summing up the results over $i$ yields

$$\sum_{i=1}^{I} \sum_{n=1}^{N} \pi_n w_{i,n}'(0) = s \cdot \left(\vec{u_k^*} - \vec{u_j^*}\right) \cdot \Lambda \cdot \left(\vec{z''} - \vec{z'}\right), \tag{A.7}$$

where $\vec{u_n^*} = \left(u_1\left(w_{1,n}^*\right), \cdots, u_I\left(w_{I,n}^*\right)\right)$ for $n = k, j$ and $\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_I \end{bmatrix}$. Notice that the left-hand side of (A.7) equals the rate of change in the compensation cost.

Now consider the perturbation that switches $A''$ and $\tilde{A}$. Similar algebraic manip-

ulation yields

$$\sum_{i=1}^{I}\sum_{n=1}^{N}\pi_n w'_{i,n}(0) = -(1-s)\cdot\left(\vec{u_k^*}-\vec{u_j^*}\right)\cdot\Lambda\cdot\left(\vec{z''}-\vec{z'}\right). \tag{A.8}$$

Since $\vec{u_k^*}-\vec{u_j^*}\neq\vec{0}$ because otherwise we can merge $A_j$ and $A_k$ together and save the monitoring cost, there are two remaining cases to consider:

(i) $\left(\vec{z''}-\vec{z'}\right)\cdot\Lambda\cdot\left(\vec{u_k^*}-\vec{u_j^*}\right)\neq 0$. In this case, (A.7) and (A.8) have the opposite signs, and the remainder of the proof follows that of Theorem 2.

(ii) $\left(\vec{z''}-\vec{z'}\right)\cdot\Lambda\cdot\left(\vec{u_k^*}-\vec{u_j^*}\right)=0$. In this case, we can always find $B'\subset A'$, $B''\subset A''$ and $\tilde{B}\subset\tilde{A}$ such that (1) $P_1(B')=P_1(B'')=P_1\left(\tilde{B}\right)<\epsilon$, (2) $\vec{z}\left(\tilde{B}\right)=(1-s')\cdot\vec{z}(B')+s'\cdot\vec{z}(B'')$ for some $s'\in(0,1)$, and (3) $(\vec{z}(B'')-\vec{z}(B'))\cdot\Lambda\cdot\left(\vec{u_k^*}-\vec{u_j^*}\right)\neq 0$. Replacing $A'$, $A''$ and $\tilde{A}$ with $B'$, $B''$ and $\tilde{B}$ in the above argument and the result follows.

□

Proof of Lemma 2.

*Proof.* Under Assumption 4, any cell $A=A_1\times\cdots\times A_I\in\Sigma$ satisfies

$$z_i(A)=1-\frac{\int_A p_{a_i=0,a_{-i}=1}(\vec{\omega})d\vec{\omega}}{\int_A p_{\vec{1}}(\vec{\omega})d\vec{\omega}}=1-\frac{\int_{A_i}p_{a_i=0}(\omega_i)d\omega_i}{\int_{A_i}p_{a_i=1}(\omega_i)d\omega_i},$$

where the right-hand side depends only on $A_i$. Plugging this into Part (i) of Theorem 4 and the result follows. □

Proof of Lemma 3

*Proof.* First, the monitoring technology of any individual incentive contract that induces the high effort from all agents has at least $2^I$ cells. Second, there exists a bi-partitional monitoring technology, e.g., $\{\{\vec{z}:\sum_i a_i z_i\geq\kappa,\forall I\},\{\vec{z}:\sum_i a_i z_i<\kappa\}\}$, that induces the high effort from all agents when equipped with the optimal wage scheme. □

Proof of Proposition 2.

*Proof.* Part (i): see the original proof of Holmstrom (1982).

Part (ii): for each $N \geq 2$, define

$$W_N = \inf_{\mathcal{P}:|\mathcal{P}|\leq N} \sum_{i=1}^{I} \sum_{A\in\mathcal{P}} w_i^*(A;\mathcal{P})$$

as the infimum of the minimum compensation costs that are attained by monitoring technologies with at most $N$ cells. Notice that $W_N$ is well-defined for all $N \geq 2$ and is non-increasing in $N$. Hence there exists $\underline{\mu} > 0$ such that $W_N + \mu \cdot f(N) > W_{2^I-1} + \mu \cdot f\left(2^I - 1\right)$ for all $N \geq 2^I$ and $\mu > \underline{\mu}$. Thus for all $\mu > \underline{\mu}$, any $\mathcal{P}^*(\mu)$ has at most $2^I - 1$ cells and hence is a group monitoring technology. □

# B Online Appendix (For Online Publication Only)

## B.1 Existence of Optimal Incentive Contract

In the baseline model, suppose the range of $Z$ is compact and connected in $\mathbb{R}$.

**Assumption 5.** *$Z(\Omega)$ is compact and connected in $\mathbb{R}$.*

**Theorem 5.** *Under Assumptions 1 and 5, the solution to Problem (2.1) exists in the following situations:*

*(i)* $H(\mathcal{P}, a) = f(|\mathcal{P}|)$, *where* $f : \mathbb{N} \to \mathbb{R}$ *is increasing and unbounded above;*

*(ii)* $H(\mathcal{P}, a) = \begin{cases} h\left(\vec{\pi}(\mathcal{P}, a)\right) & \text{if } |\mathcal{P}| \leq K \\ +\infty & \text{if } |\mathcal{P}| > K \end{cases}$, *where* $K \in \mathbb{N} - \{1\}$ *and* $h : \Delta^K \to \mathbb{R}$ *is continuous.*

*Proof.* Part (i): suppose for simplicity and for now that the distribution of $Z$ given $a = 1$ is atomless. From Theorem 2, it follows that any $\mathcal{P}^* = \{A_1, \cdots, A_N\}$ can be characterized by $N + 1$ cutoff $z$-values $\hat{z}_0 \leq \cdots \leq \hat{z}_N$ where $A_n = \{\omega : Z(\omega) \in [\hat{z}_{n-1}, \hat{z}_n]\}$ for all $n = 1, \cdots, N$. Hence the principal's problem can be solved in two steps:

Step 1. For each $K \geq 2$, find $K + 1$ cutoff $z$-values $\hat{z}_0 \leq \cdots \leq \hat{z}_K$ that mimimize the compensation cost. Define

$$\mathcal{Z}_K = \left\{(\hat{z}_0, \cdots, \hat{z}_K) : \hat{z}_n \in Z(\Omega), n = 0, \cdots, K \text{ and } \hat{z}_0 \leq \cdots \leq \hat{z}_K\right\},$$

and equip $\mathcal{Z}_K$ with the sup-norm. Let $\vec{\hat{z}} = (\hat{z}_0, \cdots, \hat{z}_K)$, and write $W\left(\vec{\hat{z}}\right)$ instead of $W(\mathcal{P})$ (the minimum compensation cost under $\mathcal{P}$). Formalize the principal's problem as follows:

$$\min_{\vec{\hat{z}} \in \mathcal{Z}_K} W\left(\vec{\hat{z}}\right).$$

Since $W$ is continuous in its argument and $\mathcal{Z}_K$ is compact, the solution to this problem exists. Denote a solution by $\vec{\hat{z}}^*$ and let $W_K = W\left(\vec{\hat{z}}^*\right)$ be the minimal compensation cost attained by $\vec{\hat{z}}^*$.

Step 2. Solve for $\min_{K \geq 2} W_K + \mu \cdot f(K)$. Since $W_K$ is non-increasing in $K$ whereas $f(K)$ is increasing and unbounded above, the solution to this problem exists.

Now suppose the distribution of $Z$ given $a = 1$ has atoms. Then the principal's choice variable becomes $((\hat{z}_0, p_0), \cdots, (\hat{z}_K, p_K))$, where $\hat{z}_n \in Z(\Omega)$ is a cutoff $z$-value, and $p_n \in [0, 1]$ denote the probability of assigning $\{\omega : Z(\omega) = \hat{z}_n\}$ to $A_n$. Plugging this to the argument above and the result follows.

Part (ii): for simplicity, consider only the case where the distribution of $Z$ given $a = 1$ is atomless. For any profile $\vec{\hat{z}} \in \mathcal{Z}_K$ of cutoff $z$-values, write the monitoring cost as a function of $\vec{\hat{z}}$, i.e., $h\left(\vec{\hat{z}}\right) = h\left(P_1(Z(\omega) \in [\hat{z}_0, \hat{z}_1]), \cdots, P_1(Z(\omega) \in [\hat{z}_{K-1}, \hat{z}_K])\right)$, and notice that $h\left(\vec{\hat{z}}\right)$ is continuous in its argument. Formalize the principal's problem as follows:

$$\min_{\vec{\hat{z}} \in \mathcal{Z}_K} W\left(\vec{\hat{z}}\right) + \mu \cdot h\left(\vec{\hat{z}}\right).$$

Since $\mathcal{Z}_K$ is compact and the objective function is continuous, the solution to this problem exists. □

## B.2 Individual Rationality

In the baseline model, suppose the agent faces an outside option that confers a reservation utility $\underline{u}$ at the contracting stage. An incentive contract $\langle \mathcal{P}, w(\cdot) \rangle$ constitutes a pair of monitoring technology $\mathcal{P}$ and wage scheme $w : \mathcal{P} \to \mathbb{R}$. It is individually rational if

$$\sum_{A \in \mathcal{P}} P_1(A) u(w(A)) \geq c + \underline{u}. \tag{IR}$$

The optimal incentive contract minimizes the total implementation cost, subject to (IC) and (IR).

**Corollary 3.** *Under Assumption 1, any $\mathcal{P}^*$ can be expressed as $\{A_1, \cdots, A_N\}$ where*

*(i) $z(A_1) < \cdots < z(A_N)$ and $w^*(A_1) < \cdots < w^*(A_N)$;*

*(ii) Each $A_n \in \mathcal{P}^*$ is Z-convex.*

*Proof.* Part (i): take $\mathcal{P}^* = \{A_1, \cdots, A_N\}$ as given and reduce the principal's problem to

$$\begin{aligned} &\min_{\{w_n\}} \sum_{n=1}^{N} \pi_n w_n, \\ &\sum_{n=1}^{N} \pi_n u(w_n) z_n \geq c, \qquad \text{(IC)} \\ &\sum_{n=1}^{N} \pi_n u(w_n) \geq c + \underline{u}. \qquad \text{(IR)} \end{aligned}$$

Let $\lambda$ and $\mu$ denote the Lagrange multiplier associated with the (IC) constraint and the (IR) constraint, respectively. In the principal's problem, taking derivative with respect to $w_n$ yields

$$u'(w_n^*) = \frac{1}{\lambda z_n + \mu}.$$

Hence if $z_j = z_k$ for any $A_j, A_k \in \mathcal{P}^*$, then merging $A_j$ and $A_k$ into a single cell has no effect on (IC) or (IR) but saves the monitoring cost, a contradiction. Thus $\mathcal{P}^*$ can be written as $\{A_1, \cdots, A_N\}$ where $z(A_1) < \cdots < z(A_N)$ and $w^*(A_1) < \cdots < w^*(A_N)$.

Part (ii): suppose, to the contrary, that some $A_j \in \mathcal{P}^*$ is not $Z$-convex. First, consider the perturbation in the proof of Theorem 2 that switches $A'$ and $\tilde{A}$, where

$$\begin{cases} z_j'(0) = \dfrac{s\,(z'' - z')}{\pi_j} \\ z_k'(0) = -\dfrac{s\,(z'' - z')}{\pi_k}, \\ z_n'(0) = 0, \forall n \neq j, k. \end{cases}$$

Take any profile $(w_n(\epsilon))_{n=1}^N$ of wages that satisfies (IC) and (IR) after this perturbation, i.e.,

$$\lambda \cdot \sum_{n=1}^{N} \pi_n u(w_n(\epsilon)) z_n(\epsilon) = \lambda \cdot c,$$
$$\text{and } \mu \cdot \sum_{n=1}^{N} \pi_n u(w_n(\epsilon)) = \mu \cdot (c + \underline{u}).$$

In these new conditions, taking total derivative with respect to $\epsilon$ and summing up the results yields

$$\sum_{n=1}^{N} \pi_n \cdot u'\left(w_n^*\right) \cdot (\lambda z_n + \mu) \cdot w_n'(0) = -\lambda \left[u\left(w_j^*\right) \cdot \pi_j z_j'(0) + u\left(w_k^*\right) \cdot \pi_k z_k'(0)\right].$$

Now since $u'\left(w_n^*\right) = \frac{1}{\lambda z_n + \mu}$ for all $n = 1, \cdots, N$, it follows that

$$u'\left(w_n^*\right) \cdot (\lambda z_n + \mu) \cdot w_n'(0) = w_n'(0), \forall n = 1, \cdots, N.$$

Substituting this into the previous condition yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = s\left[u\left(w_k^*\right) - u\left(w_j^*\right)\right](\lambda z'' - \lambda z'), \tag{B.1}$$

where the left-hand side equals the rate of change in the expected wage.

Now consider the perturbation that switches $\tilde{A}$ and $A''$. Similar algebraic manipulation yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = -(1-s)\left[u\left(w_k^*\right) - u\left(w_j^*\right)\right](\lambda z'' - \lambda z'). \tag{B.2}$$

Since $u\left(w_j^*\right) \neq u\left(w_k^*\right)$, the right-hand sides of (B.1) and (B.2) have the opposite signs. The rest of the proof follows that of Theorem 2. □

## B.3 Cost sharing

In the baseline model, suppose a fraction $\beta \in [0, 1]$ of the monitoring cost is borne by the agent.

**Corollary 4.** *Under Assumption 1, for any $\beta \in [0,1]$, Theorems 1 and 2 hold if*

*(i)* $H(\mathcal{P}, a) = f(|\mathcal{P}|)$;

*(ii)* $H(\mathcal{P}, a) = h(\vec{\pi}(\mathcal{P}, a))$ *for some totally differentiable function* $h$, *and*

$$u\left(w^*\left(A_n\right)\right) + \beta \frac{\partial h}{\partial \pi_n}\left(P_0\left(A_1\right), \cdots, P_0\left(A_N\right)\right)$$

*differs across all* $n = 1, \cdots, N$ *under the optimal contract.*

*Proof.* Part (i): same as the proofs of Theorem 1 and 2.

Part (ii): in this case, the agent's incentive compatibility constraint becomes

$$\sum_{n=1}^{N} \pi_n u(w_n) z_n - c \geq \beta \left[h\left(\pi_1, \cdots, \pi_N\right) - h\left(\pi_1 z_1, \cdots, \pi_N z_N\right)\right] \tag{IC$_\beta$}$$

Suppose, to the contrary, that some cell of $\mathcal{P}^*$ is not $Z$-convex. First, consider the perturbation in the proof of Theorem 2 that switches $A'$ and $\tilde{A}$. Take any profile of wages $\{w_n(\epsilon)\}_{n=1}^N$ such that (1) $w_1(\epsilon) = w_1(0) = 0$ and (2) (IC$_\beta$) holds after this perturbation, i.e.,

$$\sum_{n=1}^{N} \pi_n u(w_n(\epsilon)) z_n(\epsilon) - c = \beta \left[h\left(\pi_1, \cdots, \pi_N\right) - h\left(\pi_1 z_1(\epsilon), \cdots, \pi_N z_N(\epsilon)\right)\right].$$

Similar algebraic manipulation as that in the proof of Theorem 2 yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = s[v_k - v_j]\left(\lambda z'' - \lambda z'\right), \tag{B.3}$$

where

$$v_n = u\left(w_k^*\right) + \beta \frac{\partial h}{\partial \pi_k}(P_0(A_1), \cdots, P_0(A_N)), n = k, j.$$

Next, consider the perturbation in the proof of Theorem 2 that switches $\tilde{A}$ and $A''$. Similar algebraic manipulation yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = -(1-s)[v_k - v_j]\left(\lambda z'' - \lambda z'\right). \tag{B.4}$$

Since $v_k \neq v_j$, the right-hand sides of (B.3) and (B.4) have the opposite signs. The remainder of the proof follows that of Theorem 2. □

## B.4 Random Monitoring Technology

**Setup** In the baseline model, let the monitoring technology $\mathcal{P} : \Omega \to \Delta^N$ map each $\omega \in \Omega$ to a finite probability vector $\mathcal{P}(\omega) = (p_1(\omega), \cdots, p_N(\omega)) \in \Delta^N$, and $w : \{1, \cdots, N\} \to \mathbb{R}_+$ map each monitoring outcome $n \in \{1, \cdots, N\}$ to a non-negative wage $w_n \geq 0$. Time evolves as follows:

1. The principal commits to an incentive contract $\langle \mathcal{P}, w \rangle$;
2. The agent privately exerts an effort $a \in \mathcal{A} = \{0, 1\}$;
3. Nature draws $\omega \in \Omega$ according to $P_a$;
4. The monitoring technology publicly announces $n \in \{1, \cdots, N\}$ with probability $p_n(\omega)$;
5. The principal pays the promised wage $w_n$.

In this setting, $(\mathcal{P}, a)$ induces a random performance measure $X : \Omega \to \{1, \cdots, N\}$, where

$$\pi_n(\mathcal{P}, a) \triangleq P_X(X = n \mid a) = \int p_n(\omega) dP_a(\omega).$$

Let $\vec{\pi}(\mathcal{P}, a) = (\pi_n(\mathcal{P}, a))_{n=1}^N$ denote the probability vector that $(\mathcal{P}, a)$ induces. Define $Z : \Omega \to \mathbb{R}$ and $z(A)$ for each $A \in \Sigma$ the same as before. For each $n \in \{1, \cdots, N\}$, define the $z$-value of monitoring outcome $n$ by

$$z_n(\mathcal{P}) = \frac{\int p_n(\omega) Z(\omega) dP_1(\omega)}{\int p_n(\omega) dP_1(\omega)}.$$

A contract is incentive compatible if

$$\sum_{n=1}^{N} \pi_n(\mathcal{P}, 1) z_n(\mathcal{P}) u(w_n) \geq c.$$

At any given level $a$ of agent effort, the total cost of implementing an incentive contract is

$$\sum_{n=1}^{N} \pi_n(\mathcal{P}, a) w_n + \mu \cdot h\left(\vec{\pi}(\mathcal{P}, a)\right),$$

where $h$ satisfies Assumption 1. An optimal incentive contract minimizes the total implementation cost, subject to the agent's incentive compability and limited liability constraints, i.e.,

$$\min_{N \in \mathbb{N}, \langle \mathcal{P}, w \rangle} \sum_{n=1}^{N} \pi_n(\mathcal{P}, 1) w_n + \mu \cdot h\left(\vec{\pi}(\mathcal{P}, 1)\right), \text{ s.t. (IC) and (LL).}$$

**Analysis** The next definition generalizes $Z$-convexity.

**Definition 9.** *$\mathcal{P}$ is $\vec{u}$-*inbetween *for some $\vec{u} \in \Delta^N$ if for any $\omega', \omega'', \tilde{\omega} \in \Omega$ whereby $Z(\tilde{\omega}) = (1-s) \cdot Z(\omega') + s \cdot Z(\omega'')$ for some $s \in (0,1)$, we have*

$$\min\left\{\mathcal{P}(\omega') \cdot \vec{u}, \mathcal{P}(\omega'') \cdot \vec{u}\right\} \leq \mathcal{P}(\tilde{\omega}) \cdot \vec{u} \leq \max\left\{\mathcal{P}(\omega') \cdot \vec{u}, \mathcal{P}(\omega'') \cdot \vec{u}\right\}.$$

**Theorem 6.** *Under Assumption 1, any $\langle \mathcal{P}^*, w^* \rangle$ satisfies the following properties:*

*(i) $z_1(\mathcal{P}^*) < z_2^*(\mathcal{P}^*) < \cdots < z_N(\mathcal{P}^*)$ and $w_1^* = 0 < w_2^* < \cdots < w_N^*$;*

*(ii) $\mathcal{P}^*$ is $(u(w_1^*), \cdots, u(w_N^*))$-inbetween.*

*Proof.* For brevity, write $z_n$ and $\pi_n$ instead of $z_n(\mathcal{P})$ and $\pi_n(\mathcal{P}, 1)$.

Part (i): take $\mathcal{P}^*$ as given and reduce the principal's problem to the following:

$$\begin{aligned} &\min_{\{w_n\}} \sum_{n=1}^{N} \pi_n w_n, \\ &\text{s.t. } \sum_{n=1}^{N} \pi_n u(w_n) z_n \geq c, && \text{(IC)} \\ &\text{and } w_n \geq 0, n = 1, \cdots, N. && \text{(LL)} \end{aligned}$$

Taking derivative with respect to $w_n$ yields

$$u'(w_n^*) = 1/\max\{\lambda z_n, \hat{z}\}.$$

Thus if $w_j^* = w_k^*$ for some $j \neq k$, then merging $j$ and $k$ into a single monitoring outcome has no incentive effect but saves the monitoring cost. Furthermore, $w_n^* = 0$ for some $n$ whereby $z_n < 0$, because $\sum_{n=1}^N \pi_n z_n = 0$. Thus we can rank the monitoring outcomes according to their $z$-values such that $z_1^* < z_2^* < \cdots < z_N^*$ and $w_1^* = 0 < w_2^* < \cdots < w_N^*$.

Part (ii): take any $A', A'', \tilde{A} \in \Sigma$ such that

(1) $P_1(A') = P_1(A'') = P_1\left(\tilde{A}\right) = \epsilon$ for some small $\epsilon > 0$;

(2) $z\left(\tilde{A}\right) = (1-s)z(A') + sz(A'')$ for some $s \in (0,1)$.

For each $n = 1, \cdots, N$ and $A \in \left\{A', A'', \tilde{A}\right\}$, define

$$p_n(A) = \frac{\int_A p_n(\omega) dP_1(\omega)}{\int_A dP_1(\omega)}.$$

For brevity, write $p'_n = p_n(A')$, $p''_n = p_n(A'')$, $\tilde{p}_n = p_n\left(\tilde{A}\right)$, $\tilde{z} = z\left(\tilde{A}\right)$, $z' = z(A')$ and $z'' = z(A'')$.

Consider two perturbations to $\mathcal{P}^*$:

(a) Apply $(p'_1, \cdots, p'_N)$ to $\omega \in \tilde{A}$ and $(\tilde{p}_1, \cdots, \tilde{p}_N)$ to $\omega \in A'$;

(b) Apply $(p''_1, \cdots, p''_N)$ to $\omega \in \tilde{A}$ and $(\tilde{p}_1, \cdots, \tilde{p}_N)$ to $\omega \in A''$.

Under Assumption 1, neither perturbation affects the monitoring cost. We now evaluate their effects on the compensation cost.

Consider perturbation (a) first. Specifically, let $(z_n(\epsilon))_{n=1}^N$ denote the $z$-values of the monitoring outcomes after this perturbation. Straightforward algebra shows that

$$z'_n(0) = \frac{1}{\pi_n}\left(\tilde{p}_n - p'_n\right)(z' - \tilde{z}), \forall n = 1, \cdots, N.$$

Take any profile $(w_n(\epsilon))_{n=1}^N$ of wages (which clearly exists) that satisfies two conditions: (1) $w_1(\epsilon) = w_1(0) = 0$, and (2) (IC) holds after perturbation (a), i.e.,

$$\sum_{n=1}^N \pi_n u(w_n(\epsilon)) z_n(\epsilon) = c.$$

In this new (IC) constraint, taking total derivative with respect to $\epsilon$ and multiplying the result by $\lambda$ (the Lagrange multiplier associated with the (IC) constraint prior to the perturbation) yields

$$\begin{aligned}\sum_{n=1}^{N} \pi_n \cdot u'\left(w_n^*\right) \cdot \lambda z_n \cdot w_n'(0) &= -\lambda \sum_{n=1}^{N} u(w_n^*) \cdot \pi_n z_n'(0) \\ &= s \cdot \vec{u^*} \cdot \left(\vec{\tilde{p}} - \vec{p'}\right) \cdot \left(\lambda z'' - \lambda z'\right).\end{aligned}$$

where $\vec{u^*} = \left(u\left(w_1^*\right), \cdots, u\left(w_N^*\right)\right)$, $\vec{\tilde{p}} = \left(\tilde{p}_1, \cdots, \tilde{p}_N\right)$ and $\vec{p'} = \left(p_1', \cdots, p_N'\right)$. Furthermore, since $u'\left(w_n^*\right) = \frac{1}{\lambda z_n}$ for all $n \geq 2$ and $w_1'(0) = 0$, it follows that

$$u'\left(w_n^*\right) \cdot \lambda z_n \cdot w_n'(0) = w_n'(0), \forall n = 1, \cdots, N.$$

Plugging in this into the previous condition yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = s \cdot \vec{u^*} \cdot \left(\vec{\tilde{p}} - \vec{p'}\right) \cdot \left(\lambda z'' - \lambda z'\right), \tag{B.5}$$

where the left-hand side gives the rate of change in the compensation cost.

Now consider perturbation (b). Similar algebraic manipulation yields

$$\sum_{n=1}^{N} \pi_n w_n'(0) = -(1-s) \cdot \vec{u^*} \cdot \left(\vec{\tilde{p}} - \vec{p''}\right) \cdot \left(\lambda z'' - \lambda z'\right). \tag{B.6}$$

Since the right-hand sides of these conditions have the same signs, it follows that $\min\left\{\vec{p'} \cdot \vec{u^*}, \vec{p''} \cdot \vec{u^*}\right\} \leq \vec{\tilde{p}} \cdot \vec{u^*} \leq \max\left\{\vec{p'} \cdot \vec{u^*}, \vec{p''} \cdot \vec{u^*}\right\}$. □

# References


Baker, G. P. and T. N. Hubbard (2004): "Contractibility and Asset Ownership: On-Board Computers and Governance in U. S. Trucking," *Quarterly Journal of Economics*, 119(4), 1443-1479.

Banker, R., and S. Datar (1989): "Sensitivity, Precision and Linear Aggregation of Signals for Performance Evaluation," *Journal of Accounting Research*, 27(1), 21-39.

Baiman, S., and J. Demski (1980): "Economically Optimal Performance Evaluation and Control Systems," *Journal of Accounting Research*, 18(S), 184-220.

Bloom, N., and J. Van Reenen (2006): "Measuring and Explaining Management Practices Across Firms and Countries," *Centre for Economic Performance Discussion Paper*, No. 716.

——— (2007): "Measuring and Explaining Management Practices Across Firms and Countries," *Quarterly Journal of Economics*, 122(4): 1351-1408.

——— (2010): "Why Do Management Practices Differ Across Countries?," *Journal of Economic Perspectives*, 24(1), 203-224.

Bloom, N., R. Sadun, and J. Van Reenen (2012): "Americans Do IT Better: U.S. Multinationals and the Productivity Miracle," *American Economic Review*, 102(1), 167-201.

Blumrosen, L., N. Nisan, and I. Segal (2007): "Auctions with Severely Bounded Communication," *Journal of Artificial Intelligence Research*, 28, 233-266.

Boning, B., C. Ichniowski and K. L. Shaw (2007): "Opportunity Counts: Teams and the Effectiveness of Production Incentives," *Journal of Labor Economics*, 25(4), 613-650.

Bracken, D., C. Timmreck, and A. Church (2001): *The Handbook of Multisource Feedback*. San Francisco, CA: Jossey-Bass.

Bresnahan, T., E. Brynjolfsson and L. Hitt (2002): "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *Quarterly Journal of Economics*, 117(1), 339-376.

Bryson, A., R. Freeman, C. Lucifora, M. Pellizzari, and V. Perotin (2011): "Paying for Performance: Incentive Pay Schemes and Employees' Financial Participation," in *Executive Remuneration and Employee Performance-Related Pay: A Transatlantic Perspective,* ed. by T. Boeri, C. Lucifora and K. J. Murphy. Oxford University Press.

Caroli, E. and J. Van Reenen (2001): "Skill-Biased Organizational Change? Evidence from A Panel of British and French Establishments," *Quarterly Journal of Economics*, 116(4), 1449-1492.

Cover, T. M., and J. A. Thomas (2006): *Elements of Information Theory.* Hoboken, NJ: John Wiley & Sons, 2nd ed.

Chassang, S. (2010): "Building Routines: Learning, Cooperation, and the Dynamics of Incomplete Relational Contracts," *American Economic Review*, 100(1), 448-465.

Crémer J., L. Garicano, and A. Prat (2007): "Language and the Theory of the Firm," *Quarterly Journal of Economics*, 122(1), 373-407.

Dessein, W., A. Galeotti, and T. Santos (2016): "Rational Inattention and Organizational Focus," *American Economic Review*, 106(6), 1522-1536.

Dye, R. (1985): "Costly Contract Contingencies," *International Economic Review*, 26(1), 233-250.

——— (1986): "Optimal Monitoring Policies in Agencies," *The Rand Journal of Economics*, 17(3), 339-350.

Ewen, A., and M. Edwards (2001): "Readiness for Multisource Feedback," in *The Handbook of Multisource Feedback,* ed. by D. Bracken, C. W. Timmreck and A. H. Church. San Francisco, CA: Jossey-Bass.

Gibbons, R., and R. Henderson (2012): "What Do Managers Do? Exploring Persistent Performance Differences among Seemingly Similar Enterprises," in *Handbook of Organizational Economics*, Princeton, NJ: Princeton University Press.

Green, J. R. and J. J. Laffont (1987): "Limited Communication and Incentive Compatibility," in *Information, Incentives, and Economics Mechanisms: Essays in Honor of Leonid Hurwicz*, ed. by T. Groves, R. Radner and S. Reiter, Chapter 17, pp. 308-329. University of Minnesota Press.

Green, J. R. and N. L. Stokey (1983): "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, 91(3), 349-364.

Griff, R., and A. Neely (2009): "Performance Pay and Managerial Experience in Multitask Teams: Evidence from within a Firm," *Journal of Labor Economics*, 27(1), 49-82.

Halac, M., and A. Prat (2014): "Managerial Attention and Worker Engagement," *Working Paper*, Columbia University.

Hamilton, B. H., J. A. Nickerson and H. Owan (2003): "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy*, 111(3), 465-497.

Holmstrom, B. (1979): "Moral Hazard and Observability," *The Bell Journal of Economics*, 10(1), 74-91.

——— (1982): "Moral Hazard in Teams," *The Bell Journal of Economics*, 13(2), 324-340.

Holmstrom, B., and P. Milgrom (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7(S), 24-52.

Hook, B., A. Jenkins, and M. Foot (2011): *Introducing Human Resource Management*. Pearson, 6th ed.

Jensen, M. (2001): "Value Maximization, Stakeholder Theory, and the Corporate Objective Function," *Journal of Applied Corporate Finance*, 14(3), 8-21.

Kaplan, R., and D. Norton (1992): "The Balanced Scorecard: Measures that Drive Performance," *Harvard Business Review*, January-February, 71-79.

——— (1993): "Putting Balanced Scorecard to Work," *Harvard Business Review*, September-October, 134-147.

Lazear, E., and S. Rosen (1981): "Rank-Order Tournaments as Optimal Labor Contracts," *Journal of Political Economy*, 89(5), 841-864.

Li, A. (2015): "Efficiency in Dynamic Principal-Agent Models with Moral Hazard," *Working Paper*, Washington University in St. Louis.

Li, J., and N. Matouschek (2013): "Managing Conflicts in Relational Contracts," *American Economic Review*, 103(6), 2328-2351.

Maćkowiak, B., and M. Wiederholt (2009): "Optimal Sticky Prices under Rational Inattention," *American Economic Review*, 99(3), 769-803.

Madarasz, K., and A. Prat (2016): "Sellers with Misspecified Models," *Review of Economic Studies*, forthcoming.

Matějka, F., and A. McKay (2012): "Simple Market Equilibria with Rationally Inattentive Consumers," *American Economic Review: Papers and Proceedings*, 102(3), 24-29.

Milgrom, P. (1981): "Good News and Bad News: Representation Theorems and Applications," *The Bell Journal of Economics*, 12(2), 380-391.

Milgrom, P., and J. Roberts (1992): *Economics, Organization and Management,* Englewood Cliffs, NJ: Prentice Hall.

Mookherjee, D. (1984): "Optimal Incentive Schemes with Many Agents," *Review of Economic Studies*, LI, 433-446.

Nalebuff, B. J., and J. E. Stiglitz (1983): "Prizes and Incentives: Towards a General Theory of Compensation and Competition," *The Bell Journal of Economics*, 14(1), 21-43.

Pulakos, E. D. (2004): *Performance Management: A Roadmap for Developing, Implementing and Evaluating Performance Management Systems*. The Society of Human Resource Management Foundation.

Segal, I. (2006): "Lecture Notes in Contract Theory," *Unpublished Manuscript*, Stanford University.

Sims, C. (2002): "Implications of Rational Inattention," *Working Paper*, Princeton University.

——— (2006): "Rational Inattention: A Research Agenda," *Working Paper*, Princeton University.

Sobel, J. (2015): "Broad Terms and Organizational Codes," *Working Paper*, UCSD.

Solman, P. (2013): "Employee Retention: Cloud-Based Systems Enable Performance Management," *Financial Times*, November 5.

Straz, M. (2015): "Why You Need to Embrace the Big Data Trend in HR," *Entrepreneur*, April 6.

Woodley, M. (2013): "The Evolving Role of Data in Decision-Making," *The Economist Intelligence Unit*, August 12.

"The Data-ification of HR: 10 Startups Bringing Big Data to Recruitment & Talent Management," *CB Insights*, March 2015. Retrieved from www.cbinsights.com/blog/hr-big-data-startups/.