Monitoring in Multiagent Organizations*

TIM BALDENIUS, *Columbia University*

NAHUM D. MELUMAD, *Columbia University*

AMIR ZIV, *Columbia University and Interdisciplinary Center Herzliya, Israel (IDC)*

Abstract

This paper studies how to assign “monitors” to productive agents in order to generate signals about the agents’ performance that are most useful from a contracting perspective. We show that if signals generated by the same monitor are negatively (positively) correlated, then the optimal monitoring assignment will be “focused” (“dispersed”). This holds because dispersed monitoring allows the firm to better utilize relative performance evaluation. On the other hand, if each monitor communicates only an aggregated signal to the principal, then focused monitoring is always optimal since aggregation undermines relative performance evaluation.

We also study team-based compensation and randomized monitoring assignments. In particular, we show that the firm can gain from randomizing the monitoring assignment, compared with the optimal linear deterministic contract. Furthermore, under randomization, the conditional expected utility for the agent is higher when the agent is not monitored compared with the case where the agent is monitored. That is, the chance of being monitored serves as a “stick” rather than a “carrot”.

Keywords Aggregation; Monitoring; Organization design; Principal–agent; Randomization

Condensé

L’attribution des responsabilités est une tâche très importante dans les organisations décentralisées. Les chercheurs qui se sont penchés sur la théorie de la délégation en comptabilité se sont donc intéressés de près à la nature des contrats que devrait passer le mandant

* Accepted by Jerry Feltham. This paper was presented at the 2000 *Contemporary Accounting Research* Conference, generously supported by the **CGA-Canada Research Foundation**, the **Canadian Institute of Chartered Accountants**, the **Society of Management Accountants of Canada**, the **Certified General Accountants of British Columbia**, the **Certified Management Accountants Society of British Columbia**, and the **Institute of Chartered Accountants of British Columbia**. We thank two anonymous referees, Peter Ove Christensen, Srikant Datar, Jerry Feltham, Trevor Harris, Chandra Kanodia, Doron Nissim, Stefan Reichelstein, and seminar participants at the 2000 CAR Conference, City University of New York — Baruch, Columbia University (Burton Conference), Harvard University, Interdisciplinary Center Herzliya, Israel (IDC), Northwestern University, Penn State University, Rutgers University, Stanford University (Summer Camp), University of British Columbia, University of California at Berkeley, and University of Pennsylvania for helpful comments and suggestions.

d'une société, compte tenu de l'information comptable (et de l'information relative au marché) dont il dispose, avec les mandataires productifs. Jusqu'à maintenant, les chercheurs se sont toutefois moins attardés à la façon de produire ces indicateurs comptables pouvant intervenir dans les décisions contractuelles. Cette constatation étonne, puisque la production d'information destinée à l'évaluation de la performance est généralement considérée comme un objectif essentiel de tout système de contrôle de gestion. Les auteurs visent ici à mieux définir le rôle du système comptable en s'interrogeant sur la façon optimale de répartir entre un nombre limité de « contrôleurs » (ou de superviseurs non stratégiques) la responsabilité de superviser un groupe de mandataires. En d'autres termes, se demandent-ils, comment l'entreprise doit structurer son système de contrôle de gestion pour qu'il produise des indicateurs qui la renseignent le mieux possible sur la performance de ses mandataires ?

Les auteurs proposent l'exemple d'une organisation qui produit une multitude de biens et de services qu'elle vend dans différentes régions et dont la gestion relève de directeurs de produits différents. Le travail de chaque directeur de produit a une influence stochastique sur la performance du produit en question, dans chaque région. L'organisation emploie des contrôleurs dont la capacité de surveillance est limitée. Cette situation est celle dans laquelle se trouvent maintes organisations, à court terme. Elle est illustrée à la figure 1, où deux contrôleurs exercent une surveillance sur deux directeurs de produits (les mandataires) qui veillent à l'exploitation de deux établissements et fabriquent et expédient des produits dans deux régions différentes. L'organisation doit décider s'il faut demander à chaque contrôleur d'observer respectivement un seul mandataire (par exemple, en affectant un contrôleur à chaque établissement de production) ou s'il est préférable de demander à chaque contrôleur d'observer le résultat du travail des deux mandataires dans une région donnée. Cette question s'apparente de très près à un problème classique de conception des organisations : quel doit être le principal déterminant de la structure organisationnelle — gammes de produits, lieux géographiques, domaines fonctionnels ou combinaison de ces éléments (organisation matricielle) ?

Selon les constatations des auteurs, la nature des techniques de contrôle est un déterminant de la conception organisationnelle optimale que négligent habituellement les chercheurs. Les auteurs étudient en particulier l'incidence de la corrélation entre les indicateurs et l'incidence des différentes règles d'agrégation. Ils se demandent d'abord si chaque contrôleur devrait concentrer ses observations sur quelques produits (contrôle ciblé ou *focused monitoring*), peu importe le lieu géographique où ces produits sont vendus, ou étendre ses observations à plusieurs produits (contrôle réparti ou *dispersed monitoring*), mais dans un lieu géographique circonscrit. Les auteurs montrent que si les indicateurs produits par un même contrôleur sont en corrélation négative, l'attribution optimale sera celle du contrôle ciblé, soit l'attribution basée sur la gamme de produits dans l'exemple proposé. Si, par ailleurs, les indicateurs sont en corrélation positive, l'attribution optimale sera alors celle du contrôle réparti, soit l'attribution basée sur le lieu géographique. Cela s'explique par le fait que le contrôle réparti permet à l'organisation d'utiliser l'évaluation de la performance relative avec un maximum d'efficacité (voir, par exemple, Holmstrom [1982]). Les auteurs s'intéressent également à la corrélation propre au mandataire et à la corrélation propre à la région. Ils démontrent cependant que seule la corrélation propre au contrôleur est pertinente à l'attribution des responsabilités de contrôle.

Les auteurs s'intéressent ensuite à une autre caractéristique essentielle des systèmes de contrôle de gestion : l'agrégation. S'il est vrai que le supérieur immédiat d'un employé peut posséder de l'information précise sur la performance de ce dernier, le contrôleur, quant à lui, ne soumettra en général que des indicateurs agrégés au mandant. L'agrégation des données s'explique entre autres par le fait que le mandant ne possède qu'une capacité limitée de traitement de l'information pertinente. L'agrégation peut être réalisée soit en fonction du mandataire (le rapport contenant ainsi de l'information colligée par plusieurs contrôleurs), soit en fonction du contrôleur (le rapport contenant ainsi de l'information relative à plusieurs mandataires). L'analyse des auteurs porte sur l'interdépendance entre les règles d'agrégation et l'attribution des responsabilités de contrôle. Selon le système d'agrégation en fonction du contrôleur, contrairement au système de l'information désagrégée, l'attribution optimale des responsabilités de contrôle est celle du contrôle ciblé (fondée sur la gamme de produits), peu importe la nature de la corrélation. Ce qui inspire cette affirmation est que l'agrégation réduit la précision de l'évaluation relative de la performance qui représente le principal avantage du contrôle réparti. Les auteurs démontrent également que si la technique même d'agrégation est une variable du choix, l'organisation optera toujours pour un indicateur agrégé en fonction du mandataire plutôt que pour un indicateur agrégé en fonction du contrôleur — tout au moins dans la mesure où les coûts sont égaux de part et d'autre.

L'analyse des auteurs porte sur deux autres sujets liés à la conception organisationnelle : la rémunération par équipe et les mécanismes de contrôle aléatoire. Les auteurs établissent que le mandant préfère la rémunération par équipe aux contrats individuels avec chacun des mandataires, à moins que la corrélation entre les indicateurs soit positive et suffisamment élevée, auquel cas les contrats individuels basés sur le contrôle réparti sont optimaux. Dans le contexte du modèle des auteurs, cela suppose que, si la rémunération par équipe est possible et que l'on dispose d'information désagrégée, en aucun cas l'organisation ne souhaitera établir de contrat individuel avec chacun des mandataires et recourir au contrôle ciblé concurremment. Pour comprendre ce choix, il faut savoir que le contrôle ciblé domine le contrôle réparti dès que les indicateurs sont en corrélation négative, parce qu'il exploite les erreurs compensatoires de manière plus efficace. La rémunération par équipe rassemble même encore davantage d'indicateurs et permet donc au mandant de maintenir la motivation des mandataires (réfractaires au risque) à déployer des efforts, tout en allégeant le risque qu'ils doivent assumer.

Enfin, les auteurs tentent de déterminer si l'organisation peut bénéficier de l'application aléatoire du contrôle (où une probabilité est associée au contrôle des mandataires), comparativement à l'utilisation du contrat linéaire déterministe optimal. Les recherches précédentes ont démontré que l'application aléatoire du contrôle dans les *contrats*, selon une technique de contrôle déterministe, n'offre aucune valeur dans le contexte du contrat optimal lorsque les conditions sont plutôt faibles. Les auteurs envisagent une *technique de contrôle* aléatoire selon laquelle l'information livrée par l'indicateur peut être qualifiée de « très » riche à inexistante, et ils comparent ladite technique à une technique déterministe produisant un indicateur de « qualité moyenne ». Dans le cadre de référence du contrat linéaire et des mandataires multiples que choisissent les auteurs, le mandant préfère attribuer les responsabilités de contrôle de façon aléatoire. Les auteurs définissent la nature des contrats (linéaires) auxquels le mandant devrait appliquer le contrôle aléatoire et constatent que l'utilité prévue est plus élevée pour le mandataire lorsque ce dernier n'est pas soumis à un contrôle que lorsqu'il y est soumis. En d'autres termes, la possibilité du contrôle sert de « bâton » plutôt que de « carotte ».

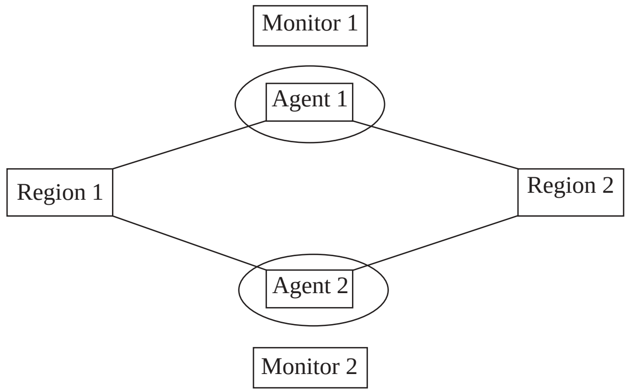
La présente étude se rattache aux travaux précédents de Holmstrom et Milgrom (1991) et de Feltham et Xie (1994) qui en sont venus à la conclusion que l'affectation des mandataires à la réalisation de différentes tâches productives améliore le contrôle, puisque la capacité de mesurer la performance varie selon les tâches. Au contraire, les auteurs s'intéressent ici plus particulièrement à la question de l'attribution optimale des responsabilités de contrôle, l'élément clé considéré étant l'interaction des techniques de contrôle, de l'agrégation de l'information et de la conception organisationnelle. Les travaux de Baiman, Larcker et Rajan (1995), de Darrough et Melumad (1995), ainsi que de Bushman, Indjejikian et Smith (1995) se rattachent, eux aussi, à la présente étude. Darrough et Melumad, de même que Baiman *et al.*, analysent la répartition optimale des responsabilités, par la société mère, entre différentes unités organisationnelles, dans le cas où la structure optimale de l'organisation dépend de la compétence de la société mère à réaliser cette tâche, par rapport à la compétence des unités organisationnelles, et de l'importance relative de l'unité organisationnelle dans la performance de la société mère. Bushman *et al.*, quant à eux, s'intéressent davantage au rôle des mesures agrégées de la performance lorsque le comportement d'un mandataire influe sur la performance des autres mandataires. Les auteurs étudient ici, par contraste, plusieurs unités organisationnelles identiques en l'absence de différences dans l'importance relative et en l'absence d'incidence externalisée directe des divisions entre elles. Ainsi peuvent-ils traiter la question de l'attribution optimale des responsabilités de contrôle.

1. Introduction

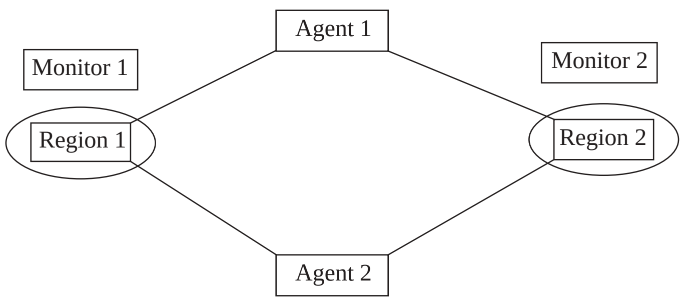
The assignment of responsibilities is a major task in decentralized organizations. Agency-theoretic research in accounting has therefore devoted significant attention to the question of how a firm's principal should contract with productive agents, based on given accounting (and market) information. How these contractible accounting signals are being generated, however, has received less attention in the literature. This is surprising given that generating information for performance evaluation is generally viewed as a major objective of a management control system. In the present paper, we aim to shed more light on the role of the accounting system by asking how the firm should optimally organize a limited number of "monitors" (non-strategic supervisors) to oversee a group of agents. Put differently, how should the firm organize its management control system for it to generate signals that are the most useful indicators of its agents' performance?[1]

Consider, for example, an organization that produces a multitude of goods and services that are sold in different regions and managed by different product managers. The effort of each product manager stochastically affects the outcome for this particular product in each region. The organization employs monitors with a limited monitoring capacity. This describes the situation faced by many organizations in the short run. Figure 1 illustrates such an organization, where two monitors oversee two product managers (agents) who operate two facilities and produce and ship products into two different regions. The firm now has to decide whether to have each monitor observe only one agent (e.g., by having the monitors reside with the production facilities), or to have each monitor observe both agents' output in a given

Figure 1 Focused (facility-based) and dispersed (region-based) monitoring



Focused (facility-based) monitoring



Dispersed (region-based) monitoring

region. This question is intimately related to a classic problem of organization design: What should be the primary determinant of organizational structure — product lines, geographical locations, functional areas, or some combination of these (i.e., matrix organization)?

Our results suggest that one determinant of the optimal organizational design, typically overlooked in the literature, is the nature of the monitoring technology. In particular, we study the impact of correlation among signals and the impact of different aggregation rules. We first address the question whether each monitor should concentrate ("focus") its observations on a few products regardless of the geographical region in which they are sold, or spread ("disperse") its observations over many products in a limited geographical region. We show that if signals generated by the same monitor are negatively correlated, then the optimal monitoring assignment will be focused — that is, product line-based in the above example. If, on the other hand, the signals are positively correlated, then optimal monitoring will be dispersed — that is, location-based. This holds because dispersed monitoring allows the firm to use relative performance evaluation most effectively (see, e.g., Holmstrom 1982). We also consider agent-specific and region-specific correlation. We show, however, that only the monitor-specific correlation is relevant for the assignment of monitoring responsibilities.

We then turn to another key feature of management control systems: aggregation. While a worker's immediate superior may possess detailed information regarding the worker's performance, the monitors will in general submit only aggregated signals to the principal. One rationale for aggregation lies in the principal's limited capacity to process all relevant information. Aggregation may be conducted either in an agent-specific way (the report thereby containing information collected by several monitors) or in a monitor-specific way (the report thereby containing information related to several agents). Our analysis focuses on the interdependence between aggregation rules and monitoring assignments. Under aggregation by monitor, in contrast to the case of disaggregate information, the optimal monitoring assignment becomes focused (or product line-based), regardless of the nature of the correlation. The intuition for this is that aggregation undermines relative performance evaluation, which is the main advantage of dispersed monitoring. We also show that if the aggregation technology itself is a choice variable, then the firm will always choose to have one aggregated signal per agent over having one aggregated signal per monitor — at least as long as these arrangements are equally costly.

Our analysis addresses two additional organizational design issues: team-based compensation and randomized monitoring arrangements. We show that the principal prefers team-based compensation over contracting with each agent individually unless the correlation between signals is positive and sufficiently high, in which case individual contracts based on a dispersed monitoring assignment are optimal. In the context of our model, this implies that, if team-based compensation is feasible and disaggregated information is available, then the firm will never want to contract with each agent individually and use focused monitoring at the same time. To understand this result, note that focused monitoring dominates dispersed monitoring whenever signals are negatively correlated, because it exploits the offsetting errors more effectively. Team-based compensation pools even more signals and, hence, allows the principal to maintain the (risk-averse) agents' incentives to exert effort while imposing less risk on the agents.

Finally, we investigate whether the firm can gain from randomizing the monitoring assignment (i.e., agents are monitored with some probability), compared with the optimal linear deterministic contract. Prior literature has demonstrated that randomization of *contracts*, based on a deterministic monitoring technology, is of no value in an optimal contracting setting under rather weak conditions.[2] We consider a randomized *monitoring technology* where the randomization falls between a "highly" informative and an uninformative signal, and compare it with a deterministic technology generating a "medium-quality" signal. In our linear contracting framework with multiple agents, the principal prefers to randomize the monitoring assignment. We characterize the nature of the (linear) contracts that the principal should randomize over, and find that the expected utility for the agent is higher when the agent is not monitored compared with the case where the agent is monitored. That is, the chance of being monitored serves as a "stick" rather than a "carrot".

The present paper is related to studies by Holmstrom and Milgrom 1991 and Feltham and Xie 1994. Both of these articles consider the allocation of productive tasks among agents to improve control, given that the ability to measure performance varies for different tasks. In contrast, we focus on the problem of optimally allocating monitoring activities, where the key consideration is the interplay of the monitoring technology, information aggregation, and organizational design. Also related to our paper are the studies by Baiman, Larcker, and Rajan 1995, Darrough and Melumad 1995, and Bushman, Indjejikian, and Smith 1995. Darrough and Melumad and Baiman et al. analyze the optimal parent company's allocation of tasks to different business units, where the optimal organization structure is a function of the parent company's task expertise relative to that of the business units and of the relative importance of the business unit to the performance of the parent firm. Bushman et al. focus on the role of aggregate performance measures when one agent's action affects the performance of other agents. In contrast to these studies, we consider multiple identical business units where there are no differences in relative importance and no direct externalities across divisions. This enables us to address the issue of optimal monitoring assignment.

The paper is organized as follows. Section 2 presents the model. In Section 3, we solve for the optimal monitoring assignment and consider the impact of aggregation. Section 4 introduces team-based compensation and randomization. In section 5, we offer our conclusions. The Appendix contains all proofs.

2. The model

The firm in our model consists of n identical productive agents and m identical monitors. The monitors were hired by the principal so as to oversee the agents. Each of the m monitors has a capacity of collecting k signals about the agents' productivity. The principal is facing the problem of completing a project of exogenously given size for which an n-dimensional effort vector, $\boldsymbol{a} = (a_1, \dots, a_n)$, is required from the agents. Hence, the problem becomes one of inducing a given effort vector at minimum cost.[3] To simplify the presentation, we focus on the case of two agents and two monitors with two signals each (i.e., $n = m = k = 2$) in the main body of the paper, and defer the solution to the more general model to the Appendix. Furthermore, we restrict attention to the case where the principal requires both identical agents to exert the same effort level — that is, $a_1 = a_2 = a$.

Each agent i, $i = 1, 2$, is assumed to be risk-averse and increasingly work-averse as expressed by the following utility function:

ASSUMPTION 1. $$\mathrm{U}_i(\cdot) = \mathrm{u}(s_i, a_i) = -\exp\left\{-r\left(s_i - \frac{a_i^2}{2}\right)\right\}.$$

That is, both agents have identical negative exponential utility functions with a common coefficient of risk aversion, r, where s_i denotes the monetary compensation.

Each agent has a reservation utility, $\bar{u}$, with a corresponding monetary certain equivalent, denoted by $\bar{s} = -(1/r)\ln(-\bar{u})$.

The agents' effort choices are unobservable and, hence, subject to moral hazard. The principal uses compensation schemes based on signals generated by the monitors (the accounting system) to motivate the agents to work. To highlight the trade-offs involved when assigning monitoring tasks and to keep the analysis simple, we assume that the monitoring is conducted by nonstrategic monitors, who themselves require no motivation or supervision. These monitors fully internalize the principal's objective function and always truthfully report their information. This modeling choice assumes away possible strategic interaction (collusion) between monitors and agents.[4]

Let the vector of contractible signals (observations) be denoted by $\boldsymbol{x} = (x_{11}, x_{12}, x_{21}, x_{22})$, where x_{jl} is the l-th observation of monitor j, and assume:

ASSUMPTION 2. *Signals are generated by a normal distribution:* $x_{jl} \sim \mathrm{N}(a_i, \sigma^2)$, *if the signal* x_{jl} *is used as an observation of agent i's performance.*

Assumption 2 implies that the expected value of an observation of an agent's performance equals the agent's effort choice.

We allow for correlation among signals and denote the covariance between any two observations generated by the same monitor by ρ, where $|\rho| < \sigma^2$.[5] Signals can be either positively or negatively correlated. Positive correlation may be the consequence of monitors using inaccurate (biased) measurement devices, where the direction and magnitude of the bias is unknown. Negative correlation among signals may occur when a monitor allocates some given measure among agents. If, for example, two agents jointly produce one unit of output, and the monitor tries to determine each agent's contribution, then overstating one agent's performance immediately implies understating the performance of the other — that is, the monitor generates negatively correlated signals.[6]

The density function over signals, conditional on the effort levels taken and the monitoring assignment, is denoted f($\boldsymbol{x}$, $\boldsymbol{a}$) and the covariance matrix associated with the joint-normally distributed vector $\boldsymbol{x}$ is:

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho & 0 & 0 \\ \rho & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & \rho \\ 0 & 0 & \rho & \sigma^2 \end{pmatrix} \quad (1).$$

Our final assumption restricts the feasible space of compensation contracts.

ASSUMPTION 3. *The principal offers each agent i a linear contract* $s_i = \alpha_{i0} + \boldsymbol{\alpha}_{i1} \cdot \boldsymbol{x}^{\boldsymbol{T}}$ *where* α_{i0} *is a fixed payment,* $\boldsymbol{\alpha}_{i1}$ *is a vector of bonus coefficients and the superscript T denotes a vector transpose.*

Assumption 3 allows for agent i's compensation to be contingent not only on his or her own performance but also on other agents' performance measures, thereby enabling the firm to use relative performance evaluation.

The principal's problem now is to minimize the total cost of inducing the desired effort vector, $\boldsymbol{a} = (a, a)$. Let $\boldsymbol{a}^1 \equiv (\tilde{a}, a)$ and $\boldsymbol{a}^2 \equiv (a, \tilde{a})$ denote the effort vectors characterized by agent i choosing an arbitrary effort level, $\tilde{a}$, and agent $j \neq i$ choosing the required value, a.

PROGRAM 1.

$$\text{Min}_{\{\alpha_{10},\, \boldsymbol{\alpha}_{11}, \alpha_{20},\, \boldsymbol{\alpha}_{21}\}} \text{TC} = \int (\alpha_{10} + \boldsymbol{\alpha}_{11} \cdot \boldsymbol{x}^{\boldsymbol{T}} + \alpha_{20} + \boldsymbol{\alpha}_{21} \cdot \boldsymbol{x}^{\boldsymbol{T}}) \mathrm{f}(\boldsymbol{x}, \boldsymbol{a}) d\boldsymbol{x},$$

subject to, for $i = 1, 2$:

$$\int -\exp\left\{-r\left(\alpha_{i0} + \boldsymbol{\alpha}_{i1} \cdot \boldsymbol{x}^{\boldsymbol{T}} - \frac{a^2}{2}\right)\right\} \mathrm{f}(\boldsymbol{x}, \boldsymbol{a}) d\boldsymbol{x} \geq \bar{u} \qquad \text{(IR)},$$

$$a \in \arg\max_{\tilde{a}} \int -\exp\left\{-r\left(\alpha_{i0} + \boldsymbol{\alpha}_{i1} \cdot \boldsymbol{x}^{\boldsymbol{T}} - \frac{\tilde{a}^2}{2}\right)\right\} \mathrm{f}(\boldsymbol{x}, \boldsymbol{a}^i) d\boldsymbol{x}, \text{ for all } \tilde{a} \qquad \text{(IC)}.$$

The two constraints ensure individual rationality and incentive compatibility in form of a Nash equilibrium, where agents choose their respective effort levels individually and noncooperatively.[7] We relax this assumption in Section 4, allowing for cooperative agent behavior. It is well known that Assumptions 1 through 3 allow for a "mean-variance" representation of the agents' preferences so that the incentive compatibility constraint can be rewritten as

$$a \in \arg\max_{\tilde{a}} \left\{\alpha_{i0} + \boldsymbol{\alpha}_{i1} \cdot \mathrm{E}[\boldsymbol{x}^{\boldsymbol{T}} | \boldsymbol{a}^i] - \frac{1}{2} r \mathrm{Var}(\alpha_{i0} + \boldsymbol{\alpha}_{i1} \cdot \boldsymbol{x}^{\boldsymbol{T}}) - \frac{\tilde{a}^2}{2}\right\}, \text{ for all } \tilde{a} \qquad \text{(IC)},$$

where $i = 1, 2$, and $\mathrm{E}[\cdot | \boldsymbol{a}]$ denotes the conditional expectation operator. We now address the issue of monitoring assignments, which is key to our analysis.

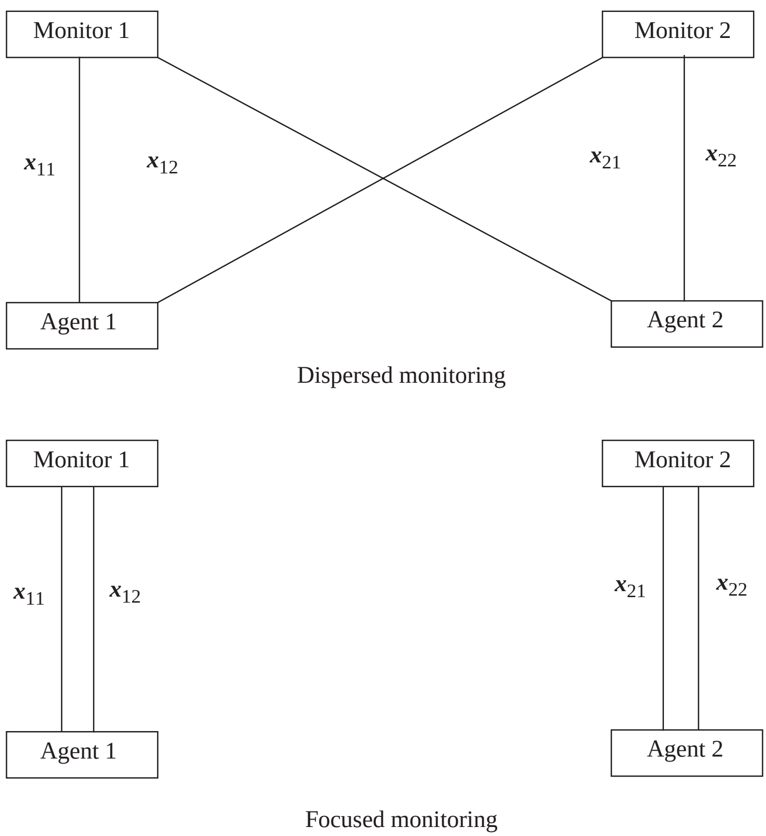
3. Optimal monitoring assignment

Contracts based on disaggregate information

We first investigate the optimal assignment of monitoring activities if compensation contracts can be based on disaggregate information, that is, if each individual component x_{jl} of the vector of signals $\boldsymbol{x}$ can be used for contracting. We differentiate between two extreme forms of monitoring assignments that we refer to as "focused" and "dispersed" monitoring. Under focused monitoring, each monitor observes only one agent so that signals x_{i1} and x_{i2} generated by monitor i are on agent i's performance. This implies that each agent is observed by only one monitor. Under dispersed monitoring, in contrast, each monitor observes both agents so that the signals x_{1i} and x_{2i} refer to agent i's performance. Figure 2 demonstrates focused monitoring (top) and dispersed monitoring (bottom) for this case.[8]

Note that the variance of an agent's compensation is decreasing and convex in the number of observations about the agent. Hence, the value to the principal of an additional observation is decreasing in the number of observations about the agent; as a result, the principal prefers an even allocation of observations among the agents.[9] If each monitor's observations are positively correlated, then the principal benefits from dispersed monitoring in two ways: (1) the interdependence among signals for each agent is minimized (in our $n = m = k = 2$ example, the observations regarding a specific agent are independent); and (2) other agents' signals can be used to discern the agent's effort (relative performance evaluation). For negative correlation, on the other hand, the first gain from dispersed monitoring becomes a disadvantage, because negatively correlated signals entail offsetting errors, which tend to favor focused monitoring. In this case, the dominant force is not immediately

Figure 2 Dispersed and focused monitoring



clear because by focusing, the principal has only a limited ability to use relative performance evaluation. However, it turns out that the first effect (offsetting errors) dominates the second effect (relative performance evaluation) for negatively correlated signals and, hence, the principal should use focused monitoring. For $\rho = 0$ (i.e., all observations are independent), any arbitrary assignment of deterministic monitoring that involves the same number of observations about each agent results in the same cost of inducing agents' effort.

PROPOSITION 1. *Suppose compensation is based on disaggregate information. Then the principal prefers* dispersed *monitoring, if the correlation among each monitor's observations is positive (i.e.,* $\rho > 0$*), and he prefers* focused *monitoring, if* $\rho < 0$.

All proofs are provided in the Appendix where we solve the model for general values of *n*, *m*, and *k*.

Using the variance terms derived in the Appendix, we find the total costs to the principal under dispersed monitoring (TC_{dis}) and focused monitoring (TC_{foc}) to be as follows:

$$TC_{dis} = 2\left[\bar{s} + \frac{a^2}{2} + \frac{ra^2}{4}\left(\sigma^2 - \frac{\rho^2}{\sigma^2}\right)\right] \quad (2),$$

$$TC_{foc} = 2\left[\bar{s} + \frac{a^2}{2} + \frac{ra^2}{4}(\sigma^2 + \rho)\right] \quad (3).$$

Comparing the two, it is obvious that dispersed monitoring involves lower costs than focused monitoring whenever

$$TC_{foc} - TC_{dis} = \frac{ra^2}{2}\rho\left(1 + \frac{\rho}{\sigma^2}\right) > 0 \quad (4).$$

Schwarz's inequality implies that $\rho < \sigma^2$; hence, (4) holds if, and only if, $\rho > 0$.

Agent-specific and region-specific correlation

So far, the term "correlation" solely referred to correlation between any two signals observed by a particular monitor. It is also conceivable that any two signals about a particular agent are correlated, even if the signals are generated by different monitors. We denote such agent-specific correlation by ρ_a. Positive agent-specific correlation, $\rho_a > 0$, may arise from unforeseen changes in the competitive environment faced by this particular product: if competition is fiercer than anticipated in all regions, then the signals will inhibit negative biases regarding the agent's performance, and vice versa for a product facing unexpectedly low competitive pressure. On the other hand, sequential observations may result in negative agent-specific correlation,

$\rho_a < 0$: if observations regarding the operating success of a division ignore accounts receivables and just capture cash receipts, then an early (late) observation tends to understate (overstate) the agent's contribution. Similarly, we also allow for region-specific correlation (e.g., unanticipated macroeconomic shocks in a certain region), and denote it by ρ_r.[10]

Recall that $\boldsymbol{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ is the vector of contractible signals, with x_{jl} denoting the l-th signal collected by monitor j. In our setting with $n = m = k = 2$, the respective covariance matrices under focused and dispersed monitoring now become:

$$\Sigma_{foc} = \begin{pmatrix} \sigma^2 & \rho + \rho_a & \rho_r & 0 \\ \rho + \rho_a & \sigma^2 & 0 & \rho_r \\ \rho_r & 0 & \sigma^2 & \rho + \rho_a \\ 0 & \rho_r & \rho + \rho_a & \sigma^2 \end{pmatrix}, \Sigma_{dis} = \begin{pmatrix} \sigma^2 & \rho + \rho_r & \rho_a & 0 \\ \rho + \rho_r & \sigma^2 & 0 & \rho_a \\ \rho_a & 0 & \sigma^2 & \rho + \rho_r \\ 0 & \rho_a & \rho + \rho_r & \sigma^2 \end{pmatrix}.$$

Note that the covariance matrix, Σ, is a function of the monitors' task allocation. Unlike monitor-specific correlation, agent-specific and region-specific correlation cannot be "filtered out" by assigning the monitors in a dispersed fashion.

PROPOSITION 1a. *Suppose $n = m = k = 2$, and compensation is based on disaggregate information. Then the principal prefers dispersed monitoring, if the monitor-specific correlation is positive (i.e., $\rho > 0$), and focused monitoring if $\rho < 0$, for all values of agent-specific and region-specific correlation (ρ_a, ρ_r).*

Proposition 1a generalizes Proposition 1 by showing that the choice between focused and dispersed monitoring is governed entirely by monitor-specific correlation. Agent-specific and region-specific correlation solely impact the *magnitude* of the difference in total costs under dispersed and focused monitoring without affecting their ranking. The total costs under dispersed and focused monitoring, respectively, are:

$$\text{TC}_{dis} = 2\left[\bar{s} + \frac{a^2}{2} + \frac{ra^2}{4}\left(\sigma^2 + \rho_a - \frac{(\rho + \rho_r)^2}{\sigma^2 + \rho_a}\right)\right] \quad (5),$$

$$\text{TC}_{foc} = 2\left[\bar{s} + \frac{a^2}{2} + \frac{ra^2}{4}\left(\sigma^2 + \rho_a + \rho - \frac{\rho_r^2}{\rho + \rho_a + \rho_r}\right)\right] \quad (6),$$

so that

$$\mathrm{TC}_{foc} - \mathrm{TC}_{dis} = \frac{ra^2}{2}\rho\left[\frac{(\sigma^2+\rho+\rho_a+\rho_r)^2}{(\sigma^2+\rho_a)(\sigma^2+\rho+\rho_a)}\right] > 0 \qquad (7),$$

which holds if and only if $\rho > 0$. For the remainder of this paper, we will therefore ignore agent-specific and region-specific correlation — that is, we assume $\rho_a = \rho_r = 0$.

Aggregation

Aggregation of information is a central feature of management control systems. In practice, supervisors submit only aggregated signals to the principal. Aggregation can be a consequence of the available information technology (inability to distinguish between individual contributions), or of a cost–benefit analysis.[11] For example, the principal may have limited information processing capacity (see Banker and Datar 1989; or Geanakoplos and Milgrom 1991). Hence, the principal can use only an aggregated number for contracting.

In this section, we consider two possible ways of aggregating information. Under *aggregation by agent*, one signal is reported per agent. Although the principal observes only one signal per agent, the principal may use (aggregated) signals regarding both agents when contracting with any given agent. Another possibility, probably less costly in terms of the collection and processing of information, is *aggregation by monitor*, whereby each monitor provides a single summary statistic for its observations. Such a case can arise when individual contributions are hard to measure.[12]

In our setting, aggregation by agent does not impose any loss because all observations gathered with regard to a given agent's output are equally weighted. Hence, the optimal allocation of monitoring tasks is identical to the one prescribed in Proposition 1. This will not be the case for aggregation by monitor because the principal receives only one aggregated signal per monitor that may relate to the performance of several agents. Our next result shows that the optimal monitoring allocation differs from the optimal allocation under no-aggregation.

PROPOSITION 2. *Suppose each monitor submits one aggregated signal. Then, the principal always prefers focused monitoring to any other monitoring assignment.*

It is instructive to compare Propositions 1 and 2 for the case of $\rho > 0$. If disaggregate information is available, then monitoring should be dispersed so as to fully utilize relative performance evaluation (Proposition 1). Under aggregation by monitor, however, the monitor-specific correlation cannot be filtered out by relative performance evaluation and, hence, monitoring should be focused even if $\rho > 0$. By focusing its observations, each monitor observes fewer agents and the aggregated signals become more informative about these agents' performance. Dispersed monitoring, in contrast, makes it harder for the principal to infer individual performance levels from aggregated signals.

Note that in our base scenario where $n = m = k = 2$, aggregation by monitor does not impose a loss on the principal under focused monitoring, because every monitor observes only one agent and, thus, aggregation by monitor corresponds to aggregation by agent.[13] For dispersed monitoring, using the cost term derived in the proof of Proposition 2, we find that

$$TC_{dis} = 2\left[\bar{s} + \frac{a^2}{2} + \frac{ra^2}{2}(\sigma^2 + \rho)\right] \quad (8),$$

which obviously exceeds the total cost under focused monitoring as given in (3).

At the organizational-design stage, different organization structures may involve different types of information aggregation. If the aggregation technology is itself a design variable, the above observations and results allow us to rank the different aggregation technologies.[14]

COROLLARY 1. *The principal prefers aggregation by agent over aggregation by monitor. If $\rho > 0$, this preference is strict.*[15]

A few remarks are in order. First, Corollary 1 may not hold in a more general model, where the monitors act strategically by misreporting information as a favor to the agents or as the outcome of side contracting with the agents. Then, if the control problem regarding the monitors' behavior is sufficiently severe, having one signal per monitor may become the preferred solution. Moreover, if the cost of aggregation by agent exceeds the cost of aggregation by monitor, the result in Corollary 1 may be reversed. Finally, it is possible to show that Proposition 2 continues to hold in the presence of both region-specific and agent-specific correlation: the principal can still attain the performance level corresponding to the no-aggregation case if only one aggregate signal per agent is available, while he always prefers focused monitoring if only one signal per monitor is available.

4. Extensions

In our above analysis, we have implicitly assumed that all agents make their effort decisions individually and non-cooperatively. Furthermore, we have assumed that the monitoring assignments — and hence the contracts offered by the principal — are deterministic. In this section, we relax both these assumptions.

Team-based compensation

First, we relax the assumption that the principal offers each agent an individual contract and that the agents choose their effort levels noncooperatively. In practice, cooperative agent behavior is a widespread phenomenon, either in the form of hidden collusion or in the form of explicit team-based compensation where the focus is on cooperation and mutual monitoring (and risk sharing) among the team members.[16] From a modeling perspective, we will not try to differentiate between these phenomena and we will simply use the term "team-based compensation" to accommodate both. If the principal chooses to contract with each agent individually, we shall maintain our earlier assumption that collusion among the agents can be prevented effectively.

Prior research has analyzed team-based compensation in organizations with a single monitor and two agents. The main findings are twofold. First, when the agents do not share any private information, then team-based compensation is weakly dominated by an incentive scheme where the principal contracts with each agent individually. Second, if the agents can monitor each others' action choices (and hence gain "collectively private" information), then team-based compensation dominates individual contracts unless the correlation between signals is positive and sufficiently high.[17] The intuition for this is that team-based compensation allows for improved monitoring and risk sharing among the agents, but it precludes relative performance evaluation. For highly positively correlated signals, the latter loss outweighs the former benefits. We integrate this finding with our results on focused versus dispersed monitoring and generalize it to a multimonitor setting.

Suppose the two agents can monitor each others' effort choices and commit to a binding side contract or, equivalently, they are compensated as a team. Denote by

$$\mathrm{s}(\boldsymbol{x}) = \beta_0 + \beta_1 \sum_{j=1}^{2} \sum_{t=1}^{2} x_{jl} \qquad (9)$$

the total compensation received by the team. Wilson (1968) has shown that the joint risk tolerance of such a team ("syndicate") equals the sum of the individual coefficients of risk tolerance. Put differently, the team's coefficient of risk aversion becomes $\frac{1}{2}r$. As a consequence, the team aims at maximizing

$$\mathrm{u}(\mathrm{s}(\cdot), a_1, a_2) = -\exp\left\{\frac{r}{2}\left(\mathrm{s}(\boldsymbol{x}) - \sum_{i=1}^{2} \frac{a_i^2}{2}\right)\right\} \qquad (10).$$

Individual rationality then requires the certain equivalent exceeds the agents' market opportunities (expressed in monetary terms):

$$\mathrm{CE}(\cdot) = \beta_0 + 2\beta_1 \sum_{i=1}^{2} a_i - \frac{r}{4}\mathrm{Var}(\mathrm{s}(\boldsymbol{x})) \geq 2\bar{s}\,.$$

Solving for the optimal contract parameters, β_0 and β_1, yields the following total cost under team-based compensation:

$$\mathrm{TC}_t = 2\left[\bar{s} + \frac{a^2}{2} + \frac{ra^2}{8}(\sigma^2 + \rho)\right] \qquad (11).$$

Comparing this last expression with (2) and (3) yields our next result.

PROPOSITION 3. *For $\rho \geq (1/2)\sigma^2$ the principal prefers to contract with each agent individually using dispersed monitoring. For $\rho < (1/2)\sigma^2$ the principal prefers team-based compensation.*

Team-based compensation is always preferred to individual contracting based on focused monitoring. This holds because a necessary condition for individual contracts to dominate team-based compensation is that $\rho > (1/2)\sigma^2$. According to Proposition 1, however, dispersed monitoring then dominates focused monitoring. If $\rho < 0$, then team-based compensation uses the resulting offsetting errors even more efficiently than do individual contracts based on focused monitoring.

Randomized monitoring

In our above analysis, we have assumed that the principal commits to a deterministic monitoring assignment (and hence to deterministic contracts) before the agents choose their effort levels. This may seem plausible given that prior research has shown that, in many cases of interest, there is no role for randomized contracts in agency relations (see, e.g., Holmstrom 1979; or Gjesdal 1982).[18] Our paper deviates from this earlier literature in that our multi-agent linear contracting setting allows for *randomized monitoring technologies*, whereas Holmstrom (1979) and Gjesdal (1982) have considered randomized contracts based on a deterministic monitoring technology in a single-agent setting. Specifically, we ask whether the principal can benefit from a management control system where monitoring assignments and compensation contracts are based on the realization of an additional random variable that is conditionally independent of the agents' actions.[19] Moreover, we aim at characterizing in more detail the specific nature of the contracts over which the principal may want to randomize.

We further simplify our basic organizational scenario by assuming that the firm employs only one monitor with a capacity of two observations to oversee two agents — that is, $n = k = 2$, $m = 1$. Slightly modifying our notation, let the vector of contractible signals now be denoted by $\boldsymbol{x} = (x_1^i, x_2^i)$, where the superscript $i = 1, 2$, associated with signal x_l^i indicates which of the two agents the l-th observation of the monitor is used on. We also ignore the issue of correlation between signals in this section by assuming that $\rho = 0$.

The monitor can observe each agent once *deterministically*, thereby using the first observation for agent 1 and the second for agent 2. Since there is no correlation among agents' signals, agent i's compensation should be independent of x_j^j, $j \neq i$: $s_i(\boldsymbol{x}) = \alpha_{i0} + \alpha_{i1} x_i^i$. Incentive compatibility requires $\alpha_{i1} = a$, which yields total costs of

$$TC_{det} = 2\left(\bar{s} + \frac{a^2}{2} + \frac{ra^2}{2}\sigma^2\right) \quad (12).$$

Alternatively, the monitor can *randomize* the allocation of observations as follows: with some probability, p_i, agent i is observed twice and paid a performance-based salary; with probability $(1 - p_i)$, he is not observed at all and paid a flat wage. In particular, let the randomized contract offered to agent i be denoted as follows:[20]

$$s_{i,\,rand}(\cdot) = \begin{cases} \beta_{i0} + \beta_{i1}(x_1^i + x_2^i), \text{ with probability } p_i \\ \gamma_i, \text{ with probability } 1 - p_i \end{cases} \quad (13).$$

The contract offered to agent j looks similar, with $p_j = 1 - p_i$. The agents' expected compensation remains linear in the vector of realized signals. However, note that the mean-variance representation for the agents' expected utility is no longer feasible in the randomized monitoring setting. The resulting distribution of outcomes can be interpreted as a lottery over normally distributed variables and, as such, it is no longer normal. While our randomization scheme may resemble an option-type contract, there is a fundamental difference in that option contracts are deterministic contracts entailing a flat and sloped range.

Consider a representative agent i. The individual rationality constraint for this agent is[21]

$$EU_i = -p_i \exp\left\{-r\left(\beta_{i0} + 2\beta_{i1}a - r\beta_{i1}^2\sigma^2 - \frac{a^2}{2}\right)\right\} - (1 - p_i)\exp\left\{-r\left(\gamma_i - \frac{a^2}{2}\right)\right\} \geq \bar{u} \quad (14).$$

Taking the first-order derivative with respect to a_i yields the following incentive compatibility constraint:

$$\beta_{i0} + 2\beta_{i1}a - r\beta_{i1}^2\sigma^2 - \gamma_i = -\frac{1}{r}\ln\left[\frac{(1 - p_i)a}{p_i(2\beta_{i1} - a)}\right] \quad (15).$$

Solving for the optimal linear compensation contract (see Appendix for details), we obtain the total cost as a function of the agents' bonus coefficients:

$$TC_{rand}(\beta_{i1}) = \sum_{i=1}^{2}\left\{\frac{a^2}{2} + p_i r\beta_{i1}^2\sigma^2 + \frac{p_i}{r}\ln\left(-\frac{2p_i\beta_{i1}}{\bar{u}a}\right) + \frac{1 - p_i}{r}\ln\left[-\frac{(1 - p_i)\beta_{i1}}{\bar{u}(2\beta_{i1} - a)}\right]\right\} \quad (16).$$

The key question for the performance comparison of deterministic versus randomized monitoring is whether there exist feasible contract coefficients $(\beta_{i0}, \beta_{i1}, \gamma_i)$ that satisfy (14) and (15) and for which $TC_{rand} \leq TC_{det}$.

PROPOSITION 4. *If compensation contracts are linear and* $\rho = 0$*, then the principal prefers randomized monitoring to deterministic monitoring.*

Proposition 4 demonstrates that the firm can improve upon the performance of the optimal linear deterministic contract by randomizing over different monitoring assignments and, thereby, over (linear) contracts.[22] While this randomized monitoring scheme still remains suboptimal, it provides a robust and very simple (and hence computationally inexpensive) improvement over a single linear contract. As such, it is of practical importance.[23] Note that, for $p = 1/2$ (the natural symmetric solution), our randomized monitoring scheme calls for the same number of *expected* observations to be allocated to each (identical) agent.

Finally, we characterize the randomized monitoring assignment in more detail. In particular, we ask whether the agent will perceive being monitored as a "carrot" or as a "stick". To address this question, we compare the certain equivalents of the agent's compensation conditional on not being observed with that conditional on being observed twice. Differentiating (16) with respect to β_{i1}, the optimal slope parameter β_{i1}^* can be shown to solve

$$2p_i r\beta_{i1}^2\sigma^2(2\beta_{i1} - a) + (2p_i\beta_{i1} - a) = 0 \qquad (17).$$

Notice that for (17) to hold, it must be that $\beta_{i1}^* \in [a/2, a/(2p_i)]$. Using $\beta_{i1}^* \leq a/(2p_i)$, we find that the right-hand side of (15) is negative and, therefore,

$$\gamma_i \geq \beta_{i0} + 2\beta_{i1}a - r\beta_{i1}^2\sigma^2 \qquad (18).$$

Recall that the left-hand side of (18) expresses the flat wage if agent *i* is not being monitored, while the right-hand side reflects the certain equivalent of his compensation if he is monitored twice. The following then is immediate.

COROLLARY 2. *Suppose that the randomized monitoring assignment is as given in (13). Then the flat wage,* γ_i*, paid to agent i if he is not monitored exceeds the certain equivalent of his expected compensation conditional on being monitored twice.*

Corollary 2 provides the fundamental intuition for Proposition 4: the chance of being monitored serves as a "stick" in that the agent's expected utility, conditional on being monitored, is less than when he is not monitored. This creates additional variance in terms of the agent's utility. The agent tries to reduce this "gap" in expected utility terms by exerting more effort, because this impacts the compensation only if he ends up being monitored. Thus, introducing this additional variance with regard to the agent's compensation in this case proves to be a relatively inexpensive means of eliciting effort.[24] Corollary 2 is consistent with the casual observation that in most audit cases the audited unit is penalized for any discrepancy or irregularity but is rarely rewarded for faring well in the audit.

Kanodia (2002), in his discussion of our paper, generalizes our results and provides further insights into the randomization issue. For any increasing, strictly concave utility function where effort is modelled as negative compensation, Kanodia shows the trade-off between the ability to better motivate the agent via increasing (decreasing) the base compensation when the agent is not (is) monitored, and the risk associated with randomly switching between the risky and the certain wage. It remains as an open question whether our results and those of Kanodia can be further generalized in an optimal contracting setting.

5. Concluding remarks

In this paper, we consider the optimal assignment of monitoring tasks in a multiagent setting. We show that when signals generated by the same monitor are positively (negatively) correlated, the principal prefers dispersed (focused) monitoring, as long as all observations are available for contracting. In terms of our organization design example presented in the introduction, under positive (negative) correlation the monitoring assignment should be location-based (product-based). Information aggregation changes this trade-off in that focused (product-based) monitoring assignment becomes the optimal solution, regardless of the nature of signal correlation. This is because aggregation undermines relative performance evaluation, which in turn was the particular strength of dispersed monitoring.

In the current setting, we consider as given the number of employees, the number of monitors, and the optimal effort of each agent. In future research, these variables could be made endogenous because the firm may adopt a different organization design for different information technologies. Another interesting extension to consider is the effect of strategic supervisors who need to be motivated and are prone to collude with other agents.

Our multiagent setting allows us to address several organizational design issues such as information aggregation, team-based compensation, and randomized monitoring technologies. We do so within the framework of linear compensation schemes, which, for tractability reasons, have been studied extensively in recent years (e.g., Holmstrom and Milgrom 1991; Bushman and Indjejikian 1993; Feltham and Xie 1994; Dutta and Reichelstein 1999). This trend has generated a debate regarding the robustness of linear contracts (see Baiman 1990; Lambert 2001). The present paper contributes to this debate by establishing that the firm should optimally randomize the monitoring assignment if it is confined to linear compensation contracts. By randomizing over linear compensation schemes, the firm can increase its expected profits without incurring the (often prohibitively high) costs of computing the optimal nonlinear contract.

Appendix: Proofs

Preliminaries: Solution to the principal's problem

We solve the principal's problem for the case of n agents, m monitors, and k signals per monitor. Denote by k_{ij} the number of observations of monitor j with regard to agent i. The only restriction on the values (n, m, k) is that they allow for symmetric monitoring assignments in the following sense: (1) all agents are observed the same number of times (requiring $\eta \equiv mk/n$ to be an integer), (2) all agents are observed by the same number of monitors, denoted by τ, and (3) all τ monitors observing a certain agent collect the same number of observations, k_{ij}, with regard to this agent. Then $\tau = \eta / k_{ij}$ holds and a monitoring assignment is uniquely determined by the parameter k_{ij} (or, equivalently, by τ).

Since all agents are identical in terms of their utility functions and required effort levels, a, they will all be offered identical contracts and we can focus attention on a representative agent, dropping the subscript i. Using this, we can rewrite Program 1:

PROGRAM 1. $\text{Min}_{\{\alpha_0, \alpha_1\}}$ $\text{TC} = n\int(\alpha_0 + \boldsymbol{\alpha}_1^{\boldsymbol{T}} \cdot \boldsymbol{x})\text{f}(\boldsymbol{x}, \boldsymbol{a})d\boldsymbol{x}$, *subject to (IR) and (IC).*

It is well known that the intercept of a linear contract solely serves the purpose of allocating the certain equivalent between the principal and the agents. Hence, one can first confine attention to a relaxed program where the (IR) constraints are ignored. Consider the (IC) constraint for the representative agent:

$$a \in \arg\max_{\tilde{a}} \left\{ \alpha_0 + \boldsymbol{\alpha}_1^{\boldsymbol{T}} \cdot \text{E}[\boldsymbol{x}|\boldsymbol{a}^i] - \frac{1}{2} r\text{Var}(\alpha_0 + \boldsymbol{\alpha}_1^{\boldsymbol{T}} \cdot \boldsymbol{x}) - \frac{\tilde{a}^2}{2} \right\}, \text{ for all } \tilde{a},$$

with $\boldsymbol{a}^i$ denoting the effort vector where all agents other than the representative one choose the required effort level, a.

Assumption 2 implies that the variance of the agent's compensation is independent of the agent's effort. Moreover, given that all monitors are identical, the vector of bonus parameters $\boldsymbol{\alpha}_1$ for the representative agent consists of only three values: α_1, α_2, and zero. Here, α_1 denotes the incentive weight attached to each observation made on the representative agent's output, while α_2 denotes the weight placed on each observation collected by any of the τ monitors, but relating to an agent other than the representative one. It is easy to show that all signals generated by the $(m - \tau)$ monitors that do not observe the representative agent at all will be given zero weight.

The agent's incentive compatibility constraint then reads:

$$a \in \arg\max_{\tilde{a}} \left\{ \tau[\alpha_0 + k_{ij}\alpha_1\tilde{a} + (k - k_{ij})\alpha_2 a] - \frac{1}{2} r\text{Var}(\text{s}(\boldsymbol{x})) - \frac{\tilde{a}^2}{2} \right\} \quad \text{(IC).}$$

Taking the first-order condition of (IC), we find that incentive compatibility requires that

$$\alpha_1 = \frac{a}{\tau k_{ij}} = \frac{a}{\eta} \quad \text{(19).}$$

The principal's problem for any monitoring allocation is to choose the bonus parameters attached to observations conducted with respect to other agents, α_2, so as to minimize the total variance of the representative agent's compensation:

$$\text{Var}(s(\boldsymbol{x})) = \tau\Big\{[k_{ij}\alpha_1^2 + (k - k_{ij})\alpha_2^2]\sigma^2 + [(k_{ij} - 1)k_{ij}\alpha_1^2 + (k - k_{ij} - 1)(k - k_{ij})\alpha_2^2 + 2k_{ij}(k - k_{ij})\alpha_1\alpha_2]\rho\Big\} \quad (20),$$

subject to the incentive compatibility constraint in (19). Setting the first-order derivative of Var(s($\boldsymbol{x}$)) with respect to α_2 equal to zero, using (19), yields

$$\tau(k - k_{ij})\alpha_2\sigma^2 + \tau(k - k_{ij} - 1)(k - k_{ij})\alpha_2\rho + (k - k_{ij})\rho a = 0 \quad (21).$$

Dividing by $(k - k_{ij})$, we get[25]

$$\alpha_2 = -\frac{\rho a}{\tau A} \quad (22),$$

where $A \equiv \sigma^2 + (k - k_{ij} - 1)\rho$.

Now we use (19) and (22) for the variance term in (20) to get

$$\begin{aligned}\text{Var}(s(\boldsymbol{x})) &= \frac{a^2}{\tau k_{ij}}\Big[\sigma^2 + k_{ij}(k - k_{ij})\frac{\rho^2\sigma^2}{A^2} + (k_{ij} - 1)\rho \\ &\quad + k_{ij}(k - k_{ij} - 1)(k - k_{ij})\frac{\rho^3}{A^2} - 2k_{ij}(k - k_{ij})\frac{\rho^2}{A}\Big] \\ &= \frac{a^2}{\eta A}[\sigma^4 + (k - k_{ij} - 1)\sigma^2\rho - k_{ij}(k - k_{ij})\rho^2 + (k_{ij} - 1)\sigma^2\rho \\ &\quad + (k_{ij} - 1)(k - k_{ij} - 1)\rho^2] \\ &= \frac{a^2}{\eta A}\{\sigma^2[\sigma^2 + (k - 2)\rho] - (k - 1)\rho^2\}.\end{aligned}$$

We note that only the denominator of Var(·) is affected by the monitoring assignment via $A \equiv \sigma^2 + (k - k_{ij} - 1)\rho$. ■

Proof of Proposition 1

Symmetry of the monitoring allocations requires that all agents are observed the same number of times — that is, $\tau k_{ij} = mk/n = \eta$. The principal's program can be restated as follows:

$$\min_{k_{ij}} \frac{ra^2\{\sigma^2[\sigma^2 + (k-2)\rho] - (k-1)\rho^2\}}{2\eta[\sigma^2 + (k - k_{ij} - 1)\rho]} \quad (23),$$

$$\text{subject to: } \exists\, \tau \in \{1, \dots, m\} \subset \mathbf{N}, \text{ such that } \tau k_{ij} = \eta. \quad (24).$$

It immediately follows that k_{ij} should take the smallest feasible value, subject to (24), if $\rho > 0$ (i.e., dispersed monitoring). The opposite holds for $\rho < 0$. ■

Proof of Proposition 1a

Let α_3 denote the coefficient in the representative agent's compensation that is attached to any of the two observations made on the other agent under focused monitoring. The variance in the representative agent's compensation then equals:

$$\text{Var}_{foc} = 2(\alpha_1^2 + \alpha_3^2)(\sigma^2 + \rho + \rho_a) + 4\alpha_1\alpha_3\rho_r.$$

Incentive compatibility requires that $\alpha_1 = a/2$. This we can use to solve for the value of α_3 that minimizes the above variance term:

$$\alpha_3 = -\frac{a\rho_r}{2(\sigma^2 + \rho + \rho_a)}.$$

As a result, the total variance in the agent's compensation under focused monitoring becomes

$$\text{Var}_{foc} = \frac{a^2}{2}\left(\sigma^2 + \rho + \rho_a - \frac{\rho_r^2}{\sigma^2 + \rho + \rho_a}\right) \quad (25).$$

We proceed in a similar fashion for dispersed monitoring where the total variance is

$$\text{Var}_{dis} = 2(\alpha_1^2 + \alpha_2^2)(\sigma^2 + \rho_a) + 4\alpha_1\alpha_2(\rho + \rho_r).$$

Again, using $\alpha_1 = a/2$, by incentive compatibility, and minimizing the variance with respect to α_2 yields

$$\alpha_2 = -\frac{a(\rho + \rho_r)}{2(\sigma^2 + \rho_a)}.$$

As a result:

$$\mathrm{Var}_{dis} = \frac{a^2}{2}\left[\sigma^2 + \rho_a - \frac{(\rho + \rho_r)^2}{\sigma^2 + \rho_a}\right] \tag{26}$$

A comparison of (25) and (26) reveals that $\mathrm{Var}_{foc} \geq \mathrm{Var}_{dis}$, whenever

$$\rho \frac{(\sigma^2 + \rho + \rho_a + \rho_r)^2}{(\sigma^2 + \rho_a)(\sigma^2 + \rho + \rho_a)} \geq 0.$$

By positive semi-definiteness of the covariance matrix, this last condition is satisfied, if and only if $\rho \geq 0$. ■

Proof of Proposition 2

Aggregation by monitor implies that for every monitor j that observes the representative agent at least once, it must be that $\alpha_1 = \alpha_2 = \alpha$ because the principal cannot distinguish between any two observations made by the same monitor, even if these observations were made for agents other than the representative one. The representative agent's compensation is then

$$\mathrm{s}(\boldsymbol{x}) = \alpha_0 + \alpha \sum_{j \in M} \sum_{l=1}^{k} x_{jl} \tag{27}$$

where M is the set of monitors that observe the representative agent. The total variance associated with $\mathrm{s}(\boldsymbol{x})$ equals

$$\mathrm{Var}(\mathrm{s}(\boldsymbol{x})) = \tau\,[k\alpha^2\sigma^2 + (k-1)k\alpha^2\rho] = \frac{\eta k \alpha^2}{k_{ij}}\,[\sigma^2 + (k-1)\rho] \tag{28}$$

Consider the representative agent's first-order condition regarding his or her effort choice that amounts to $\alpha = a/\eta$. Notice that the monitoring allocation does not affect the bonus parameter α. Using $\alpha = a/\eta$ for (28), the principal's problem is to choose the monitoring assignment, as given by k_{ij}, so as to minimize

$$\mathrm{Var}(\mathrm{s}(\boldsymbol{x})) = \frac{k\alpha^2}{\eta k_{ij}}\,[\sigma^2 + (k-1)\rho],$$

subject to (24). Clearly, this problem is solved by focused monitoring — that is, setting k_{ij} equal to the highest value that satisfies (24). Note that when $\rho < 0$ and $m < n$ there is a strict loss from aggregation by monitor, even though the monitoring allocation is not changed. ■

Proof of Proposition 3

We prove Proposition 3 for the case where $k = n$ and n/m is an integer. Each agent will then be observed $\eta = mk/n = m$ times, and positive definiteness of the covariance matrix requires that $\rho \in (-\sigma^2/(n-1), \sigma^2)$. Under team-based compensation, the agents form a "syndicate" in that their joint risk tolerance equals nr^{-1} (see Wilson 1968). The team's certain equivalent is given by

$$\text{CE}(\cdot) = \beta_0 + \sum_{i=1}^{n}\left[\eta\beta_1 a_i - \frac{a_i^2}{2}\right] - \frac{r}{2n}\text{Var}(\text{s}(\boldsymbol{x})),$$

where

$$\text{Var}(\text{s}(\boldsymbol{x})) = mk\beta_1^2\sigma^2 + m(k-1)k\beta_1^2\rho = mk\beta_1^2[\sigma^2 + (k-1)\rho].$$

Thus, incentive compatibility requires that $\alpha_1 = a/\eta$, and hence the total cost under team-based compensation becomes

$$\text{TC}_t = n\left(\bar{s} + \frac{a^2}{2}\left\{1 + \frac{r}{mn}[\sigma^2 + (n-1)\rho]\right\}\right) \quad (29),$$

since, by assumption, $n = k$.

We now turn to individual contracts based on focused or dispersed monitoring. Under focused monitoring, $n = k$ implies $k_{ij} = m$, which yields total costs of

$$\text{TC}_{foc} = n\left(\bar{s} + \frac{a^2}{2}\left\{1 + \frac{r}{m}\left[\sigma^2 + \frac{(m-1)\sigma^2\rho - (n-1)\rho^2}{\sigma^2 + (n-m-1)\rho}\right]\right\}\right) \quad (30).$$

Proceeding similarly for dispersed monitoring, $n = k$ implies $k_{ij} = 1$ so that

$$\text{TC}_{dis} = n\left(\bar{s} + \frac{a^2}{2}\left\{1 + \frac{r}{m}\left[\sigma^2 - \frac{(n-1)\rho^2}{\sigma^2 + (n-2)\rho}\right]\right\}\right) \quad (31).$$

Note that all three total cost terms are continuous in ρ. Note also that total costs are obviously lowest under team-based compensation if $\rho = 0$ and, respectively, lowest under dispersed monitoring if $\rho \to \sigma^2$. Hence, in light of Proposition 1, we only need to show that $\text{TC}_{foc} > \text{TC}_t$ for all $\rho < 0$. Comparing (30) with (29), we can rewrite the difference in total costs as follows:

$$\text{TC}_{foc} - \text{TC}_t = \frac{ra^2}{2m}[\sigma^2 + (n-1)\rho]\frac{(n-1)\sigma^2 - (2n-m-1)\rho}{\sigma^2 + (n-m-1)\rho} \quad (32).$$

For this last expression to be negative (given $\rho < 0$), $\rho < -\sigma^2/(n-1)$ has to hold which conflicts with the requirement of a positive definite covariance matrix, that is, $\rho \in (-\sigma^2/(n-1), \sigma^2)$. Hence, $TC_{foc} > TC_t$ holds within the feasible range of ρ values.

Finally, it is straightforward to show that there exists a (unique) cutoff value $\rho^o = (1/2)\sigma^2 > 0$ so that $TC_{dis} < TC_t$ if and only if $\rho > \rho^o$. ■

Proof of Proposition 4

As noted in the main body of the paper, the expected total cost of motivating a representative agent to exert the required effort level a under deterministic monitoring equals

$$TC_{i,\,det} = \bar{s} + \frac{a_i^2}{2} + \frac{1}{2} r a_i^2 \sigma^2 \qquad (33).$$

Under a randomized monitoring assignment as described in (13), in contrast, the representative agent i's expected utility is as given in (14). Incentive compatibility now requires that $\partial EU_i/\partial a_i = 0$, which can be rewritten as follows:

$$\gamma_i = \beta_{i0} + 2\beta_{i1} a - r\beta_{i1}^2 \sigma^2 + \frac{1}{r} \ln\left[\frac{(1-p_i)a}{p_i(2\beta_{i1} - a)}\right] \qquad (34).$$

Individual rationality, on the other hand, implies that $EU_i = \bar{u}$. Using (34), this amounts to

$$\beta_{i0} = \frac{a^2}{2} - 2\beta_{i1} a_i + r\beta_{i1}^2 \sigma^2 - \frac{1}{r} \ln\left[-\frac{\bar{u} a_i}{2 p_i \beta_{i1}}\right],$$

which we can use to rewrite (34) as follows:

$$\gamma_i = \frac{a^2}{2} + \frac{1}{r} \ln\left[-\frac{2(1-p_i)\beta_{i1}}{\bar{u}(2\beta_{i1} - a)}\right].$$

Thus, expected compensation for agent i under randomization, as a function of p_i and β_{i1}, becomes

$$\begin{aligned} TC_{i,\,rand}(p_i, \beta_{i1}) &= p_i(\beta_{i0} + 2\beta_{i1} a) + (1-p_i)\gamma_i \\ &= \frac{a^2}{2} + p_i r \beta_{i1}^2 \sigma^2 + \frac{p_i}{r} \ln\left[-\frac{2 p_i \beta_{i1}}{\bar{u} a}\right] \\ &\quad + \frac{1-p_i}{r} \ln\left[-\frac{2(1-p_i)\beta_{i1}}{\bar{u}(2\beta_{i1} - a)}\right]. \end{aligned}$$

Evaluating this last expression at $p = 1/2$ yields

$$\mathrm{TC}_{i,\,rand}\left(p_i = \frac{1}{2}, \beta_{i1}\right) = \frac{a^2}{2} + \frac{r}{2}\beta_{i1}^2\sigma^2 + \frac{1}{2r}\ln\left[-\frac{\beta_{i1}}{\bar{u}a}\right] + \frac{1}{2r}\ln\left[-\frac{\beta_{i1}}{\bar{u}(2\beta_{i1} - a)}\right].$$

By comparison with (33), it follows that

$$\mathrm{TC}_{i,\,rand}\left(p_i = \frac{1}{2}, \beta_{i1} = a\right) = \frac{a^2}{2}(1 + r\sigma^2) - \frac{1}{r}\ln[-\bar{u}] = \mathrm{TC}_{i,\,det} \qquad (35),$$

where the slope variable β_{i1} is evaluated at the optimal slope under deterministic monitoring, $\alpha_{i1} = a$. Moreover,

$$\left.\frac{\partial TC_{i,rand}(p_i = 1/2, \beta_{i1})}{\partial \beta_{i1}}\right|_{\beta_{i1} = a} = \left.\left[r\beta_{i1}\sigma^2 + \frac{1}{2r\beta_{i1}} - \frac{a}{2r\beta_{i1}(2\beta_{i1} - a)}\right]\right|_{\beta_{i1} = a} = ra\sigma^2 > 0 \qquad (36).$$

According to (35), the firm can replicate the expected total cost under deterministic monitoring by observing each agent half the times, and by choosing the slope term β_{i1} equal to the optimal slope under deterministic monitoring. As (36) shows, this randomization scheme is suboptimal, so that, by revealed preference,

$$\mathrm{TC}^*_{i,\,rand} < \mathrm{TC}_{i,\,rand}\left(p_i = \frac{1}{2}, \beta_{i1} = a_i\right) = \mathrm{TC}_{i,\,det},$$

where $\mathrm{TC}^*_{i,\,rand}$ denotes the expected total cost under randomization, provided that β_{i1} is chosen optimally. This completes the proof of Proposition 4. ■

Endnotes

1. This approach complements the "informativeness" literature (e.g., Holmstrom 1979) that investigates the value of an additional signal.
2. See, for example, Holmstrom 1979 or Gjesdal 1982.
3. While we take as given the number of employees and monitors, as well as the required effort level of each agent, one might want to endogenize these variables in future research because the firm may adopt a different organization design for different information technologies (see Ziv 2000).
4. A situation where supervisors work under moral hazard and where the principal uses supervisors' observations in their compensation contract is analyzed in Ziv 2000.
5. See below for the implications of agent-specific (region-specific) correlation, where any two signals collected with regard to a particular agent (in a particular region) are correlated, regardless of the monitor who collects them.

6. Likewise, one may observe negatively correlated signals when tasks are performed in sequence and the relevant measure is the time each agent used to perform his task. Also, any transfer-pricing or cost-allocation mechanism generates negatively correlated profit measurements. For an analysis of the optimal correlation level among signals in a setting where correlation is a choice variable, see Rajan and Sarath 1997.
7. In fact, we note that the agents' effort choices (as described in the IC constraint) satisfy the stronger requirement of a dominant strategy equilibrium.
8. In our basic scenario with $n = m = k = 2$, focused and dispersed monitoring are the only alternative monitoring assignments. For general values of n, m, and k, hybrid cases are conceivable. As we show in the Appendix, however, these hybrid monitoring assignments are never optimal.
9. Where different agents exert different effort levels and/or the variance terms are not identical across all agents, one can show that the proportion of the total number of observations used with regard to agent i equals $a_i\sigma_i/(\Sigma_i a_i\sigma_i)$.
10. Note that the existence of both monitor-specific and agent-specific (or region-specific) correlation can allow the principal to expand the use of relative performance evaluation. Consider a more general organization where, say, monitor 1 observes only agents 1 and 2, and monitor 2 observes only agents 2 and 3, then the observations of monitor 2 with regard to agent 3 are informative about the performance of agent 1.
11. See Banker and Datar 1989 for sufficient conditions under which linear aggregation does not entail any loss to the principal in an optimal contracting setting.
12. For example, if employees are working on an assembly line where quality control is done only after several steps are completed, the accounting system may report only the total number of defects. Another example is the case where the quality control is a pass or fail test and the cost of investigating and identifying the exact source of the problem is higher than the cost of producing another unit (e.g., production of semi-conductors).
13. This will no longer hold, if $n > m$, that is, if there are more agents than monitors. In this case, all observations collected by monitor j will be given the same weight in agent i's compensation contract, regardless of whether these signals relate to agent i's output or to another agent. As a result, each agent's compensation involves additional noise.
14. A similar consideration arises when firms provide segment reporting, where segments can be defined over geographical areas or over product lines.
15. Even when the number of monitors exceeds the number of agents (and, hence, aggregation by monitor provides more signals than aggregation by agent) the principal strictly prefers aggregation by agent, if $\rho > 0$.
16. See Holmstrom 1982 and Mookherjee 1984 for a discussion of group incentives and collusion. Here, the terms "collusion" and "cooperation" exclusively refer to side contracts between agents, while we maintain our assumption that monitors are non-strategic and hence immune to collusion/cooperation.
17. See, for example, Holmstrom and Milgrom 1990, Ramakrishnan and Thakor 1991, and Itoh 1993.
18. The literature distinguishes between ex-ante and ex-post randomization of contracts (e.g., Fellingham et al. 1984, and Arya et al. 1993). Ex-ante randomization refers to randomizing after the agents have accepted the contracts but before they have chosen

their effort. In contrast, ex-post randomization refers to randomizing *after* the agents have chosen their effort, which is the situation dealt with here.

19. See Baiman and Demski 1980 and Kanodia 1985 on stochastic monitoring. There exists a large auditing and tax literature on randomized auditing strategies (e.g., Fellingham and Newman 1985). In the literature, randomization is introduced in order to economize on monitoring costs. In our setting, the total number of observations (and hence the total monitoring cost) is given and, hence, does not change once randomization is introduced.
20. As above, we can show that both signals are given the same weight in the incentive contract.
21. Similar qualitative results are obtained by simply *assuming* mean-variance preferences.
22. Like our results for team-based compensation, we conjecture that Proposition 4 continues to hold if signals are negatively correlated, $\rho \leq 0$ (due to the advantage of focused monitoring in this setting), or if a positive correlation is sufficiently small (by continuity). Only for sufficiently large positive correlation, we expect that a dispersed and deterministic monitoring assignment will dominate as it better utilizes relative performance evaluation.
23. Our Proposition 4 is related to the approach taken in Demski and Dye 1999; they show that the firm always prefers adding a quadratic term to an otherwise linear compensation contract. See also Chen and Ziv 2002 and Feltham and Wu 2001 for a comparison of alternative simple contracts.
24. Fixed costs associated with monitoring provide a complementary rationale for a randomized monitoring strategy: a monitor has to "invest" in understanding an agent's activities in order to reduce the noise in his observations. In our setting, however, the variance, σ^2, is assumed exogenous, and, hence, such fixed cost effects are not responsible for the results of Proposition 4.
25. Note that this is feasible since α_2 will be obsolete if $k = k_{ij}$ holds.

References

Arya, A., J. Fellingham, and R. Young. 1993. Preference representation and randomization in principal-agent contracts. *Economics Letters* 42 (1): 25–30.

Baiman, S. 1990. Agency research in managerial accounting: A second look. *Accounting, Organizations and Society* 15 (4): 341–71.

Baiman, S., and J. Demski. 1980. Economically optimal performance evaluation and control system. *Journal of Accounting Research* 18 (Supplement): 184–220.

Baiman, S., D. Larcker, and M. Rajan. 1995. Organizational design for business units. *Journal of Accounting Research* 33 (2): 205–30.

Banker, R., and S. Datar. 1989. Sensitivity, precision, and linear aggregation of signals for performance evaluation. *Journal of Accounting Research* 27 (1): 21–39.

Bushman, R., and R. Indjejikian. 1993. Accounting income, stock price and managerial compensation. *Journal of Accounting and Economics* 16 (1–3): 3–24.

Bushman, R., R. Indjejikian, and A. Smith. 1995. Aggregate performance measures in business unit manager compensation: The role of intrafirm interdependencies. *Journal of Accounting Research* 33 (Supplement): 101–28.

Chen, F., and A. Ziv. 2002. Marketing-operations analysis for the provision of incentives for a multitask salesforce. Working paper, Columbia University.

Darrough, M., and N. Melumad. 1995. Divisional versus company-wide performance measures: The tradeoff between screening of talent and allocation of managerial attention. *Journal of Accounting Research* 33 (Supplement): 65–93.

Demski, J., and R. Dye. 1999. Risk, return and moral hazard. *Journal of Accounting Research* 37 (1): 27–56.

Dutta, S., and S. Reichelstein. 1999. Asset valuation and performance measurement in a dynamic agency setting. *Review of Accounting Studies* 4 (3): 235–58.

Fellingham, J., Y. Kwon, and P. Newman. 1984. Ex ante randomization in agency models. *RAND Journal of Economics* 15 (2): 290–301.

Fellingham, J., and P. Newman. 1985. Strategic considerations in auditing. *Accounting Review* 60 (4): 634–50.

Feltham, G., and M. Wu. 2001. Incentive efficiency of stock versus options. *Review of Accounting Studies* 6 (1): 7–28.

Feltham, G., and D. Xie. 1994. Performance measure congruity and diversity in multi-task principal agent relations. *Accounting Review* 69 (3): 429–53.

Geanakoplos, J., and P. Milgrom. 1991. A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and International Economies* 5 (3): 205–25.

Gjesdal, F. 1982. Information and incentives. *Review of Economic Studies* 49 (3): 373–90.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10 (1): 74-91.

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13 (2): 324–40.

Holmstrom, B., and P. Milgrom. 1990. Regulating trade among agents. *Journal of Institutional and Theoretical Economics* 146 (1): 85–105.

Holmstrom, B., and P. Milgrom. 1991. Multi-task principal-agent analysis: Incentive contracts, asset ownership and job design. *Journal of Law, Economics, & Organization* 7 (Supplement): S24–52.

Itoh, H. 1993. Coalitions, incentives, and risk sharing. *Journal of Economic Theory* 60 (2): 410–27.

Kanodia, C. 1985. Stochastic monitoring and moral hazard. *Journal of Accounting Research* 23 (1): 175–93.

Kanodia, C. 2002. Discussion of “Monitoring in multiagent organizations”. *Contemporary Accounting Research* 19 (4): 513–21.

Lambert, R. 2001. Contracting theory and accounting. *Journal of Accounting and Economics* 32 (1–3): 3–87.

Mookherjee, D. 1984. Optimal incentive schemes with many agents. *Review of Economic Studies* 51 (3): 433–46.

Rajan, M., and B. Sarath. 1997. The value of correlated signals in agencies. *RAND Journal of Economics* 28 (1): 150–67.

Ramakrishnan, R., and A. Thakor. 1991. Cooperation versus competition in agency. *Journal of Law, Economics, & Organization* 7 (2): 248–83.

Wilson, R. 1968. The theory of syndicates. *Econometrica* 36 (1): 119–32.

Ziv, A. 2000. Information technology and optimal firm structure. *Journal of Accounting Research* 38 (2): 297–328.