

Dynamics of Performance Measurement Systems^{1,2}

Pascal Courty
London Business School

Gerald Marschke
University at Albany, State University of New York

March 2003

Abstract: We review the principal agent multi-tasking literature and discuss the relevance of this literature to the implementation of performance measurement in public organizations. Arguably, the most important lesson from the literature is that performance measurement may elicit dysfunctional and unintended responses, also known as gaming responses. We propose an evolutionary model of how organizations manage performance measures when gaming is revealed over time. The model stresses the dynamic nature of the performance measure selection process---a feature that has been overlooked in the literature. We present evidence from a job training program that is consistent with the model and we discuss implications for the selection of performance measures.

JEL: H72, J33, L14

Keywords: Performance Incentive, Performance Measurement, Gaming, Multitasking, Government Organization.

¹ Comments welcome at Pascal Courty, Department of Economics, London Business School, Regent's Park, NW1 4SA, London, UK, pcourty@london.edu or Gerald Marschke, Department of Economics and Department of Public Administration and Policy, BA-110, University at Albany, State University of New York, Albany, NY 12222, marschke@albany.edu. We are grateful to James Heckman and Jeff Smith for helpful comments.

² This paper was prepared for a special issue of the *Oxford Review of Economic Policy* entitled "Financing and Managing Public Services" edited by Margaret Stevens and Paul Grout.

I. INTRODUCTION

A sizeable portion of the economic literature is devoted to understanding how organizations should design performance-based contracts (Baker, 2002). Economists have also investigated both how performance measurement and compensation systems are actually designed in practice and how agents respond to these systems (Prendergast, 1999). Much of the recent policy and public administration literature is concerned with performance measurement, as interest in accountability and efficiency has waxed in recent years and government organizations that seek to implement performance measurement face the challenge of finding the right performance measures and using them appropriately (e.g. Wholey and Hatry, 1992, and Burgess, Propper, and Wilson, 2002).

This paper reviews the principal agent multi-tasking literature with a special emphasis on its implications for the design of performance measurement systems in public service organizations. We take for granted the decision to adopt performance measurement and asks how this should be done. Others have asked when performance measurement should be used as an alternative to alternative organizational designs and the interested reader should refer to these works (e.g. Tirole, 1994, and Dixit, 2002). We also assume that the designers of the performance measurement system agree on the organizational goal since our focus is on the construction of a measurement system that captures that goal. We recognize that this assumption may not apply to all public sector organizations, as some public organisations are characterised by multiple principals with possibly conflicting goals. The main interest of this review, however, is prescriptive and this justifies our approach. From a descriptive point of view, our analysis applies to situations where there is a fairly wide consensus on what the purpose of the organization should be.

While the multi-tasking model offers no set-formula for selecting performance measures and placing weights on these measures, two issues that have challenged practitioners, it does suggest useful guidelines for constructing successful measurement systems. The main point of this literature is that performance measures communicate objectives that may not exactly correspond to the organization's true goal and this misalignment may result in inefficient resource allocations. An important lesson is that incentive designers should beware that performance measures elicit dysfunctional and unintended responses because line workers acquire in their daily routine a superior understanding of how the measurement systems work, and how performance outcomes can be manipulated.

Anecdotal stories of dysfunctional responses to performance measurement abound. Because of Gosplan's policy of motivating factory managers with incentive-backed performance targets, or quotas, the former Soviet Union is a rich source of such stories. A famous editorial cartoon appearing in a popular Soviet weekly magazine and lampooning Gosplan showed factory workers producing one very large 510 ton nail to meet their 510-ton nail quota.³ While an exaggeration, the nail story reflects myriad of real-life responses to ill-conceived performance measures that are only slightly less preposterous. Consider the case of chandeliers, once mentioned in an official Soviet speech as an offending industry. Because the original production description chosen was weight, the Soviets could soon boast the heaviest chandeliers in the world (Loucks and Whitney, 1973, pp. 302-3).⁴ Soviet stores stocked with too small nails and too bright light bulbs have also been blamed on performance-driven incentive systems.

While this evidence is anecdotal and focuses on command economies, a recent literature has begun to systematically identify and measure dysfunctional responses in government and for-profit organizations. We review the literature studying dysfunctional responses in public organizations, including applications to government job training bureaucrats (Courty and Marschke 1997 and 2004, and Heckman et al. 2002), high school teachers (Jacob and Levit, 2002), health care workers (Goddard et al., 2000 and Dranove et al. 2002), and navy recruiters (Asch, 1990), to name just a few examples. These studies identify responses that are unambiguously associated with the specifics of the incentive contract and that qualify as dysfunctional because they do not further the true goal of the organization.

Arguably, there are very few occupations where measures exist that perfectly capture the goal they are supposed to communicate. Consequently most measures are likely to elicit dysfunctional responses and this is known in the theory literature as the multi-tasking problem. The idea is that the agent can contribute to several tasks or activities and the principal cannot perfectly communicate which action the agent should take for each task. Since the performance measure imperfectly communicates the worker's contribution to the organization, the worker ends up misallocating some resources, over investing in some tasks and under investing in others.

An important limitation of the multi-tasking literature is that it provides only a static view of the process of designing performance measurement systems. This literature assumes that

³ <http://www.johnwiley.com.au/highered/eco2e/macro/stud-res/add-topics/add-topics-ch20.pdf>

⁴ Polish furniture manufacturers were also evaluated by a weight-based measure of performance, resulting in some of the "heaviest furniture in the world." (Perrin, 1998, p. 368).

although organizations use imperfect performance measures, they know the relationship between these measures and the organization's objective. In other words, the designer of the performance measurement system perfectly anticipates the gaming distortions that the performance measurement system will cause.⁵ In practice, however, the designer does not know how mis-aligned a performance measure is until it is used. In fact, if the designer knew that a performance measure was corruptible, then she should not use the measure in the first place. Accumulating evidence from the real world that dysfunctional responses prompt designers to modify measurement rules and terminate some measures seems at odds with the view that the designer knows how corruptible a performance measure is before using it.

Even organisations that eventually become successful implementing performance measurement sometimes make mistakes along the way. Consider the experience of Lincoln Electric Company, which is often held up in management education as a model for implementing high powered incentives based on a piece rate system. The company once tried to measure stenographer productivity by counting the number of times the typewriter keys were operated. The company soon observed that one of its workers was earning much more than the others. In investigating the cause, the company discovered that the worker "ate her lunch at her desk, using one hand for eating purposes and the other for punching the most convenient key on the typewriter as fast as she could" (Berg and Fast, 1975). Obviously, the company then terminated this performance measure.

We present a principal agent model of how organizations manage performance measures when dysfunctional responses are revealed over time. The principal does not know when it selects a performance measure whether it will communicate the right behavior. Only over time does the principal discover the agent's responses and then uses this additional information to update and fine-tune the incentive system. We show that imperfect information on the quality of performance measures necessarily means that the mechanism for developing performance measurement systems must be an evolutionary one. The principal adapts the measurement system in response to new information that is revealed after a measure has been activated. Consistent with this evolutionary view, we discuss some evidence from a large U.S. job training program showing that the performance measurement system has been updated several times over a period of over 20 years to address unanticipated dysfunctional responses.

⁵ We define gaming as strategic responses to performance measures that increase performance outcomes and result in costly misallocation of resources (see Holmstrom and Milgrom, 1991, Baker, 1992, and Courty and Marschke, 2004).

The main implication of our model for the design of performance measurement systems is that selecting performance measures on the basis of their correlation with the organization's true objective may not always be a valid approach. In particular this selection rule will be flawed when gaming plays an important role and in these situations the selection of performance measures has to be an experimental process.

The paper is organized as follows. The next section reviews the economic literature on the construction of performance measurement systems. Section 3 introduces a simple model capturing the dynamics of performance measurement systems. Section 4 draws some implications of our analysis to the selection of performance measures. Section 5 concludes.

II. LITERATURE REVIEW

The moral hazard model of incentive design is the main tool economists have used to understand the construction of performance measurement systems and the provision of incentives. The theoretical literature on the design of incentives is reviewed in Gibbons (1997) and Prendergast (1999). Marschke (2001) and Propper and Wilson (this issue) also review this literature and draw implications for government organizations. Dixit (2002) and Burgess and Ratto (this issue) review the broader literature on organizational design and focus on issues that are specific to the public sector. These literatures are well-established and the above surveys are excellent and exhaustive. Therefore, we are selective and deliberately focus on the issues relevant to the construction of performance measurement systems.

Broadly speaking, the principal agent framework of incentive provision has made two main contributions. The early models in this literature explain how organizations construct incentives systems to motivate and reward effort through the provision of performance bonuses.⁶ These early works, however, are less relevant for most government organizations because of such organisations' reluctance or inability to introduce high-powered performance incentives. More recent works in the principal agent literature address occasions where performance measures communicate objectives that imperfectly correspond to the organization's true goal and where this misalignment results in a misallocation of resources

⁶ This literature identifies a trade off in the design of performance bonuses between effort incentives and risk. Since the seminal work of Holmstrom (1979), this trade-off has been analysed in great detail in the literature. The basic idea is that to elicit effort the principal must link the agent's compensation to a noisy performance outcome. The outcome is random and because the agent is assumed to be risk-averse, incentive-provision imposes costs. See Prendergast (2002) for a review of the empirical evidence on the trade off between incentives and risk. Because rewards are small in government bureaucracies, we will not consider risk aversion and the risk-incentive trade-off further.

(Holmstrom and Milgrom (1991), Baker (1992 and 2002), Feltham and Xie (1994), and Banker and Datar (1989)). Our review focuses on this second strand of the literature.

The moral hazard principal agent framework has emerged in response to questions that typically originated in private sector applications. Although it is sometimes assumed in the literature that measurement problems in the public and private sectors are similar, it is legitimate to question this assumption and the relevance of this literature to public sector organization. To address this issue, we dedicate a section to explain why this literature is relevant to public sector organizations, and another section to critically review the evidence from performance measurement in the public sector that is consistent with the literature.

The Principal Agent Multi-Tasking Framework

According to the moral hazard paradigm of incentive provision, a principal contracts with an agent to perform a set of tasks. Because the principal cannot contract on what she truly cares about---in the case of a public sector organisation this might be difficult-to-measure social welfare---she contracts on an imperfect measure of performance. To illustrate how this framework applies to government organizations, we use the example of a publicly funded job training organization. The government acts as the principal and “contracts” with local, training agencies (the agents) to deliver job training services to a target population. These local agencies exercise discretion over resource allocation (e.g. make enrolment and training decisions) subject to some bureaucratic rules (e.g. eligibility rules).

For simplicity, we present a simple framework very similar to the one found in Baker (1992). Each task is characterized by its type α . The nature of these tasks is context specific. In the case of job training, we might think of each job training participant or client as a task. The principal would like the agent to invest in effort but not in gaming. The principal employs an (imperfect) performance measure to guide the agent’s efforts. In the training program example, a performance measure could be the fraction of trainees who are employed upon program completion. The agent privately observes the task’s type α (e.g., the motivation, talents, and skills of the applicant) and invests $e(\alpha)$ in effort (e.g., substantive training) and $g(\alpha)$ in gaming (e.g., subsidized employment that terminates shortly after program completion).

The insights generated from the simple case of a single performance measure generalize to multiple measures. We also assume linearity as this clearly identifies the three components

of a performance measure that have received attention in the literature. That is, we assume a performance outcome is generated by

$$m(e,g;\alpha)=v(\alpha)e+w(\alpha)g+\tilde{\varepsilon}(\alpha) \quad (1).$$

1. The first component $v(\alpha)e$ captures the investments that are perfectly aligned with the principal's true objective.
2. The second component captures gaming distortions. By assumption, the principal does not value investment in gaming. Two benchmark cases are worth emphasizing. A perfectly aligned measure is such that $w(\alpha)=0$ while a pure gaming measure is such that $v(\alpha)=0$.
3. The third component captures a measurement error term. The presence of the error term will complicate the search for good performance measures and we return to this issue in section IV.

The agent chooses an effort investment, $e^*(\alpha)$ and a gaming investment $g^*(\alpha)$, the measurement error is realized $\varepsilon(\alpha)$ and this generates the realized performance outcome for task α

$$M(\alpha)=v(\alpha)e^*(\alpha)+ w(\alpha)g^*(\alpha)+\varepsilon(\alpha),$$

The performance outcome is the sum of all task outcomes

$$M=\sum_{\alpha}(v(\alpha)e^*(\alpha)+ w(\alpha)g^*(\alpha)+\varepsilon(\alpha)).$$

The principal cannot contract on effort $e(\alpha)$ or on the true objective $v(\alpha)e(\alpha)$ but can contract on the performance outcome M . In the standard moral hazard model, the agent is paid a fixed salary and also a bonus that varies with the performance outcome. Because the agent dislikes effort and the principal cannot directly monitor effort, a performance bonus is required to elicit effort. The incentive bonus is assumed linear in performance. That is, the agent's compensation is

$$\beta_0+\beta M,$$

where β is the weight on the performance measure and β_0 is a fixed payment. The realized objective for task α is,

$$V(\alpha)=v(\alpha)e^*(\alpha)+\eta(\alpha) \quad (2),$$

where $\eta(\alpha)$ is the realization of a random noise $\tilde{\eta}(\alpha)$ that is possibly correlated with $\tilde{\varepsilon}(\alpha)$.

Note that the realized performance outcome M and the realized objective V depend on the agent's actions, which in turn depend on the incentive contract.

The principal wants the agent to invest in effort but not in gaming. To understand what happens when the performance measure is imperfect, it is useful to consider the case where $Ew^2=0$. In this case, there is no gaming investment, and when the cost of effort is quadratic, e.g. $c(e)=1/2e^2$, the optimal incentive weight equals one ($\beta=1$) and the principal achieves the first best. In general, however, the performance measure is not perfect ($Ew^2\neq 0$) and gaming takes place. The existence of gaming has two implications for the construction of measurement systems. First, given a choice, the principal will prefer those measures that generate less gaming, that is, those measures that best predict the true goal of the organization. Second, the principal factors gaming into the construction of the incentive contract, that is, the principal chooses

$$\beta = \frac{\sum_{\alpha} v^2}{(\sum_{\alpha} v^2 + \sum_{\alpha} w^2)} < 1.$$

See the appendix for a formal derivation. The optimal weight is less than one because the principal anticipates that gaming will take place and accordingly reduces the weight on the performance measure.

There is much evidence that gaming matters. A branch of the accounting literature is dedicated to identifying responses from the specifics of contracts that can only be consistent with these specifics and thus are unambiguously dysfunctional. For example, several studies investigate how agents use their discretion over the timing of performance reporting to meet performance thresholds.⁷ Healy (1985) documents that managers who are compensated for meeting annual income thresholds use their discretion over the timing of income reporting to smooth their compensation across accounting years. Similarly, Oyer (1998) reports that firms' sales show more variability at the end of fiscal years---when sales persons' bonuses are computed---than in the middle.

Why is this Framework Relevant to Public Organizations?

At this point, one may argue that a literal interpretation of the principal agent model does not capture the problems faced by most government organizations because of their reluctance or inability to introduce pay for performance. According to the standard interpretation of the model, the agent supplies no effort in the absence of incentives so there is no point in using performance measurement in the absence of bonuses. This observation suggests that the

⁷ As we will argue later, an important challenge to establish gaming is to show that the responses identified really generate efficiencies. Most studies present convincing evidence that agents respond to the performance incentives in unintended ways, but it is not always clear whether these responses impose efficiency costs to the organization. One could argue that the actions identified in the timing reporting literature are just accounting manipulations, with no consequences for the allocation of productive resources.

principal agent framework may not apply to many public service organizations since explicit financial incentives are rare and even when they are used they are typically low-powered.

Our interest in the model, however, is its implications for the choice of performance measures (not bonus determination) and we argue that the insights on performance measurement also hold in a slightly different interpretation of the principal agent framework where bonuses play no role. In this alternative interpretation, the incentive problem is solved through monitoring, or some other scheme, such that the agent supplies a given amount of effort. For example, assume that $\frac{1}{2}e^2(\alpha)$ and $\frac{1}{2}g^2(\alpha)$ correspond to the amount of effort hours and gaming hours the agent invests in task α . Through monitoring, the principal can compel the agent to allocate T hours to work overall. The agent is indifferent to how she allocates T across tasks and across effort and gaming. The principal compensates the agent for each hour worked and asks the agent to maximize the performance outcome. Because the agent is indifferent over how resources are allocated across projects, she maximizes the performance outcome subject to the constraint that

$$\frac{1}{2}\sum_{\alpha}(e^2(\alpha)+g^2(\alpha))\leq T.$$

The principal faces the problem of choosing the measure that best communicates to the agent how to allocate T. Assume the true goal of the organization is well defined (e.g., to raise the lifetime labour market earnings of job training participants in the job training example) but this goal is too complex to explain and cannot be used to direct the agent's investment decisions. The principal can communicate explicit performance measures (e.g., participants' employment rate at program completion) that are easily understood but are also imperfect statements of the organization's goal.

Under the moral hazard interpretation, the principal chooses a performance measure and an incentive weight. The performance contract is used both to elicit effort and to communicate how to allocate effort across tasks. Under the alternative interpretation, the principal chooses a performance measure and monitors the agent to ensure that investment T is supplied. Increasing T increases effort and gaming but also increases the total payroll cost the same way that increasing β increases the size of the bonus paid in the standard interpretation. In this alternative interpretation, performance measures are used only to communicate how to allocate effort, not to elicit effort, and the basic assumption is that although performance measure generate gaming the overall resource allocation still improves with the adoption of performance measurement. We show in the appendix that these two interpretations of the principal agent framework generate identical predictions for the design

of performance measurement systems (i.e. choice of performance measures). In the core of the paper, we will stick to the standard interpretation as it has received the most attention in the literature and also to make the mapping between our analysis and the literature more transparent.

The translation of the second prediction of the model in the context of public organizations is that the principal will reduce the total amount of resources controlled by the agent (denoted T), in response to distortions because the marginal social return of investment decreases. An additional response to gaming is that the principal may redesign the agent's job to reduce the agent's discretion over those tasks that are most susceptible to be gamed. In the training program application, the principal may impose eligibility rules and quotas for 'hard-to-serve' populations to address the possibility that the agent may respond to labor market-based performance measures by selecting only the most attractive applicants. This captures the idea that although the principal cannot communicate to the agent what is gaming and what isn't (the principal can only communicate an imperfect performance measure), she can still limit the agent's span of control (bureaucratic discretion) and this can reduce gaming.

This alternative interpretation of the model is consistent with many instances of the use of performance measurement in the public sector. Performance measurement systems, also known in the public management literature as performance indicator schemes or performance planning, are typically used in the absence of financial incentives and the main rationales for introducing performance measures include accountability, control, feedback, and transparency (Smith, 1995 and Heinrich, 2002). These rationales recognize the fact that explicit performance measures can improve efficiency by influencing how organizations allocate resources. The reason for introducing imperfect performance measures is that in many public organizations it is very expensive to assess the impact of decision-making on the true goal of the organization. Consider our job training example. Obtaining reliable estimates of the value that job training adds to an enrollee's set of labor market skills is costly and takes time. An experimental evaluation of the largest US training program in the late 80's involving only a fraction of the local training centers took seven years to complete and, according to one estimate, cost over \$21 million (Smith, 1996).⁸

⁸ Experimental evaluations of job training programs are expensive because to obtain meaningful results large numbers of participants must be followed for many months beyond the start of training. Non-experimental evaluations are cheaper but many policy analysts believe they produce unreliable estimates of impacts because of the difficulty of finding comparison groups that are similar to treatment

Multi-tasking Evidence from the Public Sector

The multi-tasking or gaming framework captures the notion that the investment allocation that maximizes performance outcomes does not necessarily correspond to the allocation that maximizes value-added. To illustrate gaming, we consider the case of performance measurement in schools (Burgess, Croxson and Gregg, 2002). In recent years some policy analysts and public officials have advocated setting up performance measures for local school districts, possibly backed by educational subsidies as incentives. Such performance measures are based on scores from standardized tests on reading, writing, and arithmetic. Such tests do not measure the results of teaching citizenship, conflict resolution, interpersonal skills, and other skills whose development is an important aim of primary schools. Because the tests do not measure citizenship, for example, the theory predicts that teachers will neglect this skill. Instituting performance measures can cause distortions by causing agents to spend no time on some activities that are productive but not rewarded.

Several studies have shown evidence of timing responses in the public sector.⁹ Asch (1990) shows that navy recruiters who receive awards for meeting year-end recruitment quotas respond by postponing work effort to the end of the year. In the context of a training program organization, Courty and Marschke (2004) show that training program managers strategically time the reporting of their performance outcomes. They make the important distinction between responses that divert resources (e.g., agents' time) from productive activities and responses that simply reflect an accounting phenomenon. They find that the responses they identify have a negative impact on the true goal of the organization and thus conclude that these responses are more than simply an accounting phenomenon.

The evidence on unintended responses is not limited to timing responses. Goddard et al. (2000) and Dranove et al. (2002) show evidence of dysfunctional response to explicit performance measures in the health sector. Jacob and Levitt (2002) investigate teacher cheating. Some school districts allocate school budgets on the basis of schools' performance. A number of highly publicized incidents of teacher cheating have fuelled the suspicion that teachers have responded by "teaching the test" and manipulating students' grade-to-grade promotions to boost scores. However, most of this evidence is anecdotal. Jacob and Levitt propose an innovative way to measure the extent of teacher cheating that combines measures on unexpected test score fluctuations and suspicious patterns of answers for students. They

groups. See Lalonde (1986) and Heckman, Lalonde, and Smith (2001) for discussions of the relative merits of experimental and non-experimental evaluation techniques.

show that the joint distribution of these two variables should demonstrate systematic patterns if some teachers cheat and others do not.

Courty and Marschke (2003b) propose a different approach to identify gaming. They look at how the correlation between the true goal of the organization and the performance measure changes after a measure is introduced. As we will show below, the multi-tasking model predicts that this correlation should decrease after a measure is introduced. Courty and Marschke find some evidence consistent with this prediction.

Researchers who study agent responses to performance measures often find evidence of actions that raise performance outcomes, but then find it difficult to demonstrate that these actions are sub-optimal, i.e. to show that these responses lower the stated objective of the organization. The empirical challenge is establishing what the agent's value-added would have been absent the agent's actions. Consider the cream-skimming literature in job training programs that has studied enrolment responses to performance measurement in the organization created by the Job Training Partnership Act (JTPA) of 1982. JTPA was one of the first large-scale experiments with incentive-backed performance measurement in the United States. JTPA's mission was to raise the earnings ability and lower the welfare dependency of the poor. JTPA evaluated local managers' performance by their clients' labor market success (e.g. employment status) at the end of training and rewarded successful managers with small budgetary increases. Critics of JTPA's performance incentive system feared these measures would encourage managers to enrol into their programs only those participants likely to perform well on employment measures---the most "job-ready"---irrespective of how much they might gain from the program. Some studies have found evidence that program managers prefer the "job-ready," but this alone is not evidence of gaming. To demonstrate gaming, one must show that the job-ready applicants are also those who do not benefit the most from the program (Heckman et al. 1997).

Open Issues

The economic literature on performance incentive design typically assumes that although organizations use imperfect performance measures, organizations know the relationship between these measures and the organization's objective (Prendergast, 1999). In other words, the literature assumes that the principal perfectly anticipates the gaming distortions that the performance incentive system will cause. Because the principal is perfectly informed about

⁹ Propper and Wilson (this issue) review the overall impact of performance measures in the public

this relationship, she can devise and implement at the outset the performance incentive system that optimally trades off the gain from incentives (or from communication in the alternative interpretation to public sector organizations) and the loss from distortions.

We propose an alternative model where the principal does not know how aligned or misaligned a performance measure is until it is used. After the principal implements a performance measure, however, she observes the agent's responses. Any unanticipated and unintended responses will cause the principal to re-evaluate the performance measure. We should therefore observe the principal updating the incentive system as she learns about its effectiveness.

The evolutionary view of the process of designing performance measurement systems is consistent with much anecdotal evidence that performance incentive systems are sometimes aborted because they induce unexpected and dysfunctional responses. For example, Baker et al. (1994) discuss several instances of such responses including Sears Auto Center, whose compensation system paid sales people commissions based on store revenues, with bonuses for meeting sales quotas for services and products. Sears terminated this incentive system upon learning that the system was encouraging car mechanics to mislead consumers into unnecessary repair work. Incentive systems are not systematically terminated when shortcomings are identified. Most often, they are modified and some performance measures are terminated while others are introduced. In fact, an important recommendation from incentive practitioners is that performance incentive systems should be constantly monitored and updated (Kravchuk and Schack, 1996).

III. A SIMPLE MODEL OF DYNAMIC PERFORMANCE MEASUREMENT

The evidence suggests that principals do not know when they select a performance measure whether the measure will communicate the right behavior. Some measures are more corruptible than others. Only over time does the principal find out the agent's responses and then uses this additional information to update the incentive system. We present a simple model of how organizations fine-tune incentive systems to control gaming costs. To illustrate the main ideas, we start with a simple two-period model and we discuss later how the main insights generalize.

sector. We focus in this section on dysfunctional responses.

We assume that there are many imperfect performance measures. Each performance measure perfectly captures effort and also has a gaming dimension. That is, the performance outcome is generated according to

$$m_i(e, g; \alpha) = v(\alpha)e + w_i(\alpha)g$$

where i indexes the performance measure. Because our focus is on gaming, we assume no measurement noise. Gaming opportunities are measure specific. The gaming actions that increase measure i leave measure $j \neq i$ unchanged and vice versa. We assume therefore that no two measures share the same weaknesses. The costs of effort and gaming are the same across all tasks and are respectively $1/2 e^2$ and $1/2 g^2$. We assume that the agent maximizes the performance outcome in each period. In a single agent model, this implies that the agent is myopic since she does not anticipate the principal responding to her gaming actions at the end of the first period. Assuming that the agent maximises the performance outcome in each period may not require myopia in a model with multiple agents, however, since it may still be optimal for each of the forward-looking agents to maximise current performance outcomes given that all other agents do so. We will return to this assumption when we discuss the generality of the results.

We assume that there are two types of performance measures. High gaming measures are such that $E w^2(\alpha) = w^H$ while low gaming measures are such that $E w^2(\alpha) = w^L$ with $w^H > w^L$. The principal observes whether a performance measure is high or low gaming only after it has been used.¹⁰ Before using a measure, the principal does not know its type but believes that it can be high or low with equal probability.

In each period, the principal picks a performance measure and chooses its incentive weight. We assume that trying out more than one performance measure in any period is suboptimal. This is reasonable if, for example, the principal faces a cost to collecting performance outcomes. We allow the principal to change the performance measure at the beginning of the second period, however.

We follow the literature and solve for the optimal incentive weight. In the absence of distortion, the efficient weight is $\beta = 1$. Recall that in the context of a public organization a higher incentive weight means that the principal would be willing to invest more resources in the production of the performance outcome (raise T) because the marginal return on a unit of investment is high.

¹⁰ In equilibrium, the principal can infer at the end of the first period a performance measure's type from the observation of the performance outcome M because high gaming measures generate more gaming and higher performance outcomes than low gaming ones.

Switching Performance Measures

In this two period model, the principal tries a measure in the first period. If the measure proves to be of the low-gaming type, she retains the measure in the second period, otherwise she retires that measure and introduces a new one. Let $W=1/2 (w^H+w^L)$. Because the principal does not know the type of a new measure, the optimal incentive weight on this performance measure---let us call it β_1 ---is

$$\beta_1=\sum_{\alpha}v^2/(\sum_{\alpha}v^2+W).$$

In the second period, the principal observes the performance measure's type. If the performance measure is a low gaming measure, the principal keeps it and increases its weight to

$$\beta_{2,K}=\beta=\sum_{\alpha}v^2/(\sum_{\alpha}v^2+w^L).$$

If the performance measure instead proves to be of the high gaming type, the principal switches measures at the end of period 1. In the second period, the principal sets the weight on the first period performance measure to zero and, because she does not yet know the new measure's type, she sets its weight to

$$\beta_{2,S}=\sum_{\alpha}v^2/(\sum_{\alpha}v^2+W).$$

Thus, the model predicts that the weight on a performance measure should change over time in a systematic way.

Prediction 1: *The weight on a performance measure either increases over time or the performance measure is removed.*

When the principal moves from a high to a low gaming measure, the increment in value-added (profits, if the principal is a firm owner) is

$$1/2 \beta^2(w^H-w^L)$$

where w^H-w^L can be interpreted as a measure of how imperfectly informed the principal is. This suggests that major switches in performance measures are more likely to take place early in the implementation of measurement systems when the principal lacks knowledge of gaming technologies.

Performance Measure Alignment

Do performance outcomes change in a systematic way as the contract changes? To simplify the exposition, we say that the performance measure used in period 1 is performance measure 1 and if this measure is replaced by a new measure in period 2 we call that new

measure performance measure 2. Let $M_i(\alpha, t)$ denote performance outcome corresponding to performance measure i in period t . Define $M(\alpha) - V(\alpha)$ as a measure of the alignment between a measure and the true goal. Two cases must be distinguished. Consider first the case where the principal does not change the performance measure at the end of period 1. Performance measure M_1 becomes less aligned with the true objective V as the incentive weight increases since

$$M_1(\alpha, 2) - V(\alpha, 2) = \beta_{2,K} w_1^2(\alpha) > \beta_1 w_1^2(\alpha) = M_1(\alpha, 1) - V(\alpha, 1).$$

The agent does not invest in gaming activities that are specific to performance measure two in either period. This performance measure appears perfectly aligned

$$M_2(\alpha, 2) - V(\alpha, 2) = M_2(\alpha, 1) - V(\alpha, 1) = 0.$$

Consider next the case where the principal switches performance measures at the end of period 1. The performance outcome on measure 2 at the end of the first period, that is, when only measure 1 is used in the incentive contract, is

$$M_2(\alpha; 1) = \beta_1 v^2(\alpha).$$

But this is exactly equal to the principal's observed objective and $M_2(\alpha; 1) - V(\alpha; 1) = 0$.

Performance measure 2 is a perfect proxy for the principal's objective when it is not used in the incentive contract. In period 2, performance measure 2 is used in the incentive contract and it becomes noisier since

$$M_2(\alpha; 2) - V(\alpha; 2) = \beta_{2,S} w_2^2(\alpha) > 0 = M_2(\alpha; 1) - V(\alpha; 1).$$

The opposite phenomenon happens to performance one as the agent stops investing in gaming activities that are specific to that measure

$$M_1(\alpha; 1) - V(\alpha; 1) = \beta_1 w_1^2(\alpha) > 0 = M_1(\alpha; 2) - V(\alpha; 2).$$

Prediction 2: *The alignment between a performance measures and the true goal decreases as the performance measure is activated or as it is more heavily rewarded and increases as the measure is retired.*

To illustrate this prediction, consider again the Sears Auto Center case introduced earlier. Before Sears implemented its incentive scheme, Auto Center profits may have been positively correlated with the number of repair jobs. This statistical relationship may have prompted Sears officials to compensate Auto Centers on the basis of the number of repairs completed. Once Sears began paying managers bonuses for meeting service quotas, however, those service quotas became the managers' objective. It was not long before the managers had found easy ways to boost sales volume that did not also result in higher store profits. By charging customers for unneeded and unperformed repairs, store staff uncoupled the

performance measure from the store's long-term profits. Their response to the incentives drove up the value of the performance measures while driving down profits. Thus, repairs and long-term profits would not be positively correlated after Sears based pay on the number of repairs performed.

This prediction has important implications for the selection of performance measures. Before turning to these implications, we discuss how the main predictions of the model generalize.

Discussion

Our simple two-period model suggests that performance measurement systems evolve over time. In a framework with more than two periods, the dynamic we identified carries on only until the principal finds a low gaming measure. For example, if one takes the model literally, the principal would never change measures with probability half, and would change measures at least once with probability half. A more general and realistic interpretation of the model, however, suggests that the dynamic may be more complex.

Assume, for example, that it takes time to learn how to game a performance measure. In that situation, little gaming takes place early on but gaming would increase as the agent acquires experience and learns the measure-specific gaming technology. In this more general interpretation of the model, some measures are still more easily corruptible than others, but in addition, the corruptibility of a measure increases with the agent's experience. The amount of gaming that takes place in a given period, therefore, will depend on how corruptible the measure was to start with and how long the agent has had to learn how to game the measure. Although a performance measure may be very successful when it is first used, its effectiveness may decline over time, and it may be optimal at some point to replace it. The choice of switching performance measures depends on the potential gains from reducing gaming. Consider the case where the principal moves from a high to a low gaming measure and assume the incentive weight is constant. Then the increase in profits following the switch from the high to the low gaming measure is proportional to $w^H - w^L$.

Assuming that there are two periods or more also introduces the possibility that the agent behaves strategically. Our model could be adapted to permit the agent to anticipate that the principal is likely to terminate those measures that can be easily gamed but the two main predictions of the model would still hold. They would hold because it would still be optimal for the agent to take advantage of poor performance measures. The principal will be inclined to change the measure if she believes that the agent has found ways to game the existing one.

If performance measures degrade over time, as some have argued, then there will be a dynamic to measurement systems that will not necessarily stop.

Evolution of Performance Measurement: The Case of Job Training

The model predicts that measurement systems should evolve over time. To illustrate this prediction, we review some evidence from Courty and Marschke (2002) where we investigate how performance measurement has evolved over the last twenty years in JTPA (introduced in section II). We discuss several changes that suggest that the nature of the local decision-makers' responses to performance measurement was not entirely foreseen, but once observed became the basis for important modifications. This evidence is consistent with the experiential learning process emphasized in our model.

In the United States, large-scale job training for the economically disadvantaged at the federal level dates back to the Kennedy administration. In 1982, however, the structure of federal job training programs changed dramatically with the passage of the Job Training Partnership Act (Courty and Marschke, 2003a). Largely because of the perception that a command and control type of bureaucratic structure in job training had not produced results, JTPA made job training more decentralized, delegating most decision-making authority to local training managers. In addition, the program sought to make these managers more accountable by holding them to explicit standards of performance and by backing those standards with financial and other incentives. Between 1984 and 1999, JTPA was one of the largest federally funded job training programs for the economically disadvantaged. From 2000 through the present, the program has continued, but as modified under the Workforce Investment Act of 1998.

JTPA's mission was to raise the earnings ability and lower the welfare dependency of the poor. JTPA's original set of performance measures included cost measures designed to impel program managers to give efficiency concerns weight in their decision-making. Cost-based measures judged JTPA's managers by how much they spent to produce a job placement. Over time, JTPA officials came to believe that the cost measures were encouraging short run, 'quick fix'-type activities in lieu of longer activities with more training content focusing on increasing human capital. In 1992, eight years after the cost measures were first introduced, JTPA officials phased out these measures because "research and experience have shown that the use of cost standards in the awarding of incentives has had the *unintended effect* of

constraining the provision of longer-term training programs.”¹¹ The cost measures’ implementation and removal suggest an important dynamic in the construction of performance measures. At the time they formulated the cost measures, JTPA’s performance measure designers did not entirely foresee the local decision makers’ responses. Once the designers understood these measures’ effects, however, they removed the measures.

Another important change in the measurement system is the move to “follow-up” measures. JTPA’s first labor-market performance measures were measures of enrolees’ labour market status at the time of the enrolees’ exit or termination from the training program. For example, an important labour outcome measure was the *employment rate at termination*, computed as the fraction of enrolees who were employed on the date they officially completed the program. In 1988 the JTPA designers began to phase out termination-based performance measures in response to a number of studies that seemed to show that these measures, with their emphasis on the enrolee’s employment status on the last day of training, induced training agencies to emphasize job placement-oriented services that had no long-term impact on enrolees’ skills. In 1988, the JTPA designers introduced measures based on the employment state of enrolees three months after the official end of their association with the training centres---or follow-up measures--- to “[promote] effective service to participants and [assist] them to achieve long-term economic independence.”¹² The switch from termination-date-based measures to follow-up measures constituted the second important change to the JTPA measurement system and it appeared to have been prompted by learning about unintended responses.¹³

Between 1992 and 2000, the year the Workforce Investment Act (WIA) supplanted JTPA, performance measures remained largely unchanged. Performance measurement, however, further evolved under the WIA program (Heinrich, 2002). While WIA retained the general principle of basing performance measures on the labour market outcomes of enrolees, it further refined several features of the JTPA performance measurement system. First, WIA made two changes that take into account the cream-skimming evidence developed during

¹¹ Federal Register January 5, 1990 (italics ours).

¹² State of New Jersey Performance Standards Manual, PY1988-89, Division of Employment and Training, New Jersey Department of Labor, April 1990.

¹³ The circumstances surrounding other changes to the JTPA performance measurement system suggest a pattern similar to the dynamic identified in our model. For example, Federal officials have “tried-out” some of the key components of the JTPA incentive system in the job training program that preceded JTPA. When they first introduced the employment performance measures, Federal officials noticed that training centers were failing to terminate enrolees who, while no longer taking training services, were unemployed. By holding back idle, poorly performing enrolees, training centers could boost their performance scores.

JTPA. The performance measures in WIA focus on participant outcomes measured six months following placement, whereas JTPA performance measures focused on shorter-term outcomes (90 days after program completion). WIA also includes among the JTPA-style performance measures a new before-after measure of enrollees' earnings. Conceptually, the difference between an enrollee's earnings before enrolment and after termination is more similar to an earnings or employment gain---and thus more similar to the objective of job training under JTPA and WIA---than is a post-training labour outcome. Before-after measures may therefore reduce the incentive managers had under JTPA to enrol persons only because they were likely to produce high employment and wage rates the end of their training. Second, the performance measures under WIA also include a measure of "customer satisfaction" produced from post-training surveys of enrollees and their employers.

The evidence on performance measurement under JTPA and WIA demonstrates a clear evolution that took place over a period of more than 20 years. The performance measurement system is still improving as new information and fresh research increases policy makers' understanding of how the performance measurement influences decision making and shapes training outcomes.

IV. IMPLICATIONS FOR THE SELECTION OF PERFORMANCE MEASURES

An important concern of practitioners is the selection of performance measures. What does the model say about this choice? More specifically, practitioners often select measures based on how correlated they are with the true objective of the organization (Banker, Potter, and Srinivasan,2000, and Ittner and Larcker,1998). When is a correlation criterion correct?

To address this question, it is important to clarify what is meant by correlation and to define the unit of observation. Consider again the example of JTPA. The labor market and cost outcomes used to measure performance are imperfect proxies for human capital value-added, which corresponds to the true goal of JTPA. One way to compute a correlation between these performance measures and the goal is to treat each enrollee as a task. The organization measures performance outcomes at the enrollee level (e.g. labor market outcomes) and value-added can be estimated by measuring long-term earning impacts of job training net of costs, also at the enrollee level. In general, task level observations may not be available but one can still compute correlation using cross sectional data (observations of different training agencies) or time series data (observations of the same training agency over

Under JTPA, Department of Labor officials closed this "loop hole" by limiting the time an idle enrollee

time).¹⁴ In these situations, the unit of observation is likely to aggregate a number of tasks and one has to be careful, as we argue next, in interpreting what the correlation really captures.

We return to the model to illustrate the problems that arise with the use of correlations. Assume that the principal discovers at the end of the first period that the first measure is a high gaming measure and decides to replace the measure. To guide that choice the principal may wish to select the new measure on the basis of how well it predicts the true goal. For example, one way to proceed would be to compute the correlation between all candidate measures and the true goal of the organization and to select the measure with the highest correlation. For the sake of generality, assume that performance outcomes and value-added are generated according to equations (1) and (2). Assuming that $\tilde{\varepsilon}(\alpha)$ and $\tilde{\eta}(\alpha)$ are independent of $v(\alpha)e(\alpha)$, the covariance between M and a candidate measure V is

$$\text{Cov}(M,V)=\text{Var}(v(\alpha)e(\alpha))+\text{Cov}(\tilde{\varepsilon}(\alpha),\tilde{\eta}(\alpha))$$

Two problems are worth discussing. First, note the presence of the term $\text{Cov}(\tilde{\varepsilon}(\alpha),\tilde{\eta}(\alpha))$. Selecting performance measures on the basis of a correlation criteria will tend to select measures with a high $\text{Cov}(\tilde{\varepsilon}(\alpha),\tilde{\eta}(\alpha))$. The principal does not care about this covariance, however, since the error terms do not influence the agent's resource allocation (Baker, 2002). The presence of the error terms $\tilde{\varepsilon}(\alpha)$ and $\tilde{\eta}(\alpha)$ contaminates the inference. Correlations between performance measures and the true goal will bias the selection decision when the joint distributions of $\tilde{\varepsilon}(\alpha)$ and $\tilde{\eta}(\alpha)$ vary across performance measures.

Problems with the use of correlation to rank performance measures will be exacerbated if the unit of observation aggregates many tasks. Consider the training program application where the unit of observation is a training agency. If all training agents make similar investments then they will achieve the same outcome and the covariance will say nothing about the quality of a measure since the term $\text{Var}(\sum_{\alpha}v(\alpha)e(\alpha))$ vanishes and $\text{Cov}(M,V)=\text{Cov}(\sum_{\alpha}\tilde{\varepsilon}(\alpha),\sum_{\alpha}\tilde{\eta}(\alpha))$. In practice, it is important to know whether the correlation captures mostly noise or whether it is really picking up information about the agent's choices.

Second and even more importantly, correlating a non-active performance measure and the true goal reveals nothing about the gaming term $w(\alpha)g^*(\alpha)$ that will appear after the

could remain on the books to 90 days (see Barnow and Smith, 2002, and Courty and Marschke, 2002).

performance measure will be activated. Correlation is not a useful criterion for selecting performance measures because it reveals nothing about the gaming strategies available to the agent. As demonstrated in the model, the statistical relation between a measure and the true goal is endogenous. This relation depends on the performance measurement system. One can discover how good a performance measure is only after it has been used.

This idea is reminiscent of the Lucas critique and the principle is the same.¹⁵ Relations that appear empirically stable are not stable once attempts are made to use them in policy rules. Darley (1991) noted that performance measures tend to degrade after they are used, a conclusion consistent with the model. There is also some evidence from training programs supporting the hypothesis that the relationship between short term performance outcomes and long term impacts depends on whether incentives are used at the time this relationship is measured (see Courty and Marschke, 2003b).

To summarize, a selection method for performance measures that is based on how well measures predict the true objective (using correlation or other statistical tools), as is commonly used by practitioners, has important limitations. Such a method is only valid under two restrictive assumptions: (1) the covariations between the noise terms is constant across performance measures, and (2) the principal can predict the gaming actions that will take place after the measure is used. This suggests that in many situations the principal will be able to identify a good measure only after it has been used! This is consistent with our model suggesting that the selection of performance measure has to be an experimental process.

V. CONCLUSIONS

We review the principal agent multi-tasking literature and draw implications for the implementation of performance measurement in public organizations. Although the literature does not provide a set formula to select performance measures and appropriately weight them, it does provide general guidelines to manage performance measurement systems. Arguably the most important lesson is that performance measurement may elicit dysfunctional and unintended responses, also known as gaming.

¹⁴Several studies attempting to validate JTPA's performance measures exist. For example, Heckman, Heinrich, and Smith (2002) correlate the performance outcome and value-added at the "task" (enrollee) level, while Barnow (2000) perform the correlation at the "agent" (training center) level.

¹⁵ This is also analogous to Goodhart's law stating that a measure ceases to be a good measure when it becomes a target. As stated by Goodhart (2003) "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." Goodhart's law is a sociological analogue of Heisenberg's uncertainty principle in quantum mechanics stating that measuring a system usually disturbs it.

This evidence suggests that one of the main challenges in using performance measurement is to manage gaming. We propose an evolutionary model of how organizations manage performance measures when gaming is revealed over time. The model shows that the selection process of performance measures is dynamic---a feature that has been overlooked in the literature. Consistent with this evolutionary view, we discuss some evidence from a large U.S. job training program showing that the performance measurement system has been updated several times over a period of over 20 years to address unanticipated dysfunctional responses.

This evolutionary view offers a radically different way to think about the design of performance measurement systems that departs from the static view of the performance incentive literature and from the way many practitioners think about performance measurement. An implication of our analysis is that selecting performance measures on the basis of their correlation with the organization's true objective may not always be a valid approach. In particular this selection rule will be flawed when gaming plays an important role and in these situations the selection of performance measures has to be an experimental process.

APPENDIX: MORAL HAZARD WITH RISK NEUTRAL AGENT

We follow Baker 1992's model. Assume that both the agent and the principal are risk neutral. The agent's reservation utility is U_0 . The principal uses linear contracts. That is, the agent's compensation is $\beta_0 + \beta M$, where β is the weight on the performance measure and β_0 is a fixed payment. The agent's expected utility is,

$$U = \beta_0 + \sum_{\alpha} (\beta m(e(\alpha), g(\alpha); \alpha) - 1/2 e^2(\alpha) - 1/2 g^2(\alpha)).$$

The agent's investment responses given contract scheme (β_0, β) are given by the agent's first order conditions

$$\begin{aligned} e^*(\alpha, \beta) &= \beta v(\alpha), \\ g^*(\alpha, \beta) &= \beta w(\alpha). \end{aligned}$$

Given the agent's investment response, the principal chooses the payment scheme (β_0, β) that maximizes its objective subject to the constraint that the agent participates $U > U_0$. The principal's objective is the expected net return on the projects minus the expected payoff to the agent, or Π , where

$$\Pi = \sum_{\alpha} (e^*(\alpha, \beta) v(\alpha) - (\beta_0 + \beta m(e^*(\alpha, \beta), g^*(\alpha, \beta); \alpha))).$$

The designer sets β_0 to bind the agent's participation constraint. Replacing β_0 from the participation constraint and the agent's investment response from the agent's first order conditions in the principal's objective gives

$$\Pi = \sum_{\alpha} (\beta v^2(\alpha) - 1/2 \beta^2 v^2(\alpha) - 1/2 \beta^2 w^2(\alpha)) - U_0.$$

The optimal incentive weight is

$$\beta = \sum_{\alpha} v^2 / (\sum_{\alpha} v^2 + \sum_{\alpha} w^2) \leq 1.$$

The efficient investment is $e^e(\alpha) = v(\alpha)$ and $g^e(\alpha) = 0$. As long as $Ew^2 > 0$, the agent under invests in the projects

$$e^*(\alpha) < e^e(\alpha),$$

and over invest in gaming

$$g^*(\alpha) > 0.$$

The direct cost of gaming is $1/2 \sum_{\alpha} g^{*2}(\alpha) = 1/2 \beta^2 \sum_{\alpha} w^2$. There are also indirect costs of gaming because the principal reduces the incentive weight.

Under the alternative interpretation of the principal agent model, the agent maximizes

$$\sum_{\alpha} m(e(\alpha), g(\alpha); \alpha)$$

subject to the constraint that

$$1/2 \sum_{\alpha} (e^2(\alpha) + g^2(\alpha)) \leq T.$$

The optimal investments are

$$e^*(\alpha, T) = \lambda(T)v(\alpha),$$

$$g^*(\alpha, T) = \lambda(T)w(\alpha),$$

where $\lambda(T)$ is the marginal return on investment T and is increasing in T . The principal hires the agent, offers the agent an amount equal to her utility in her outside option (U_0), and monitors that the agent invests T . The principal maximizes

$$\Pi = \sum_{\alpha} (e^*(\alpha, T)v(\alpha) - T - U_0).$$

After substituting for e^* , the problem can be written as

$$\Pi = \sum_{\alpha} (\lambda v^2(\alpha) - 1/2\lambda^2 v^2(\alpha) - 1/2\lambda^2 w^2(\alpha)) - U_0,$$

which is identical to the standard principal agent problem.

REFERENCES

- Asch, B.J. (1990), 'Do Incentives Matter? The Case of Navy Recruiters', *Industrial and Labor Relations Review*, **43**, 89S-106S.
- Baker, G. P. (1992), 'Incentive Contracts and Performance Measurement', *Journal of Political Economy*, **100**(3), 598-614.
- Baker, G. P., Gibbons, R., and Murphy, K. (1994), 'Subjective Performance Measures in Optimal Incentive Contracts', *Quarterly Journal of Economics*, **109**(2), 1125-1156.
- Baker, G. P. (2002), 'Distortion and Risk in Optimal Incentive Contracts' *Journal of Human Resources*, **37**(4), 728-751.
- Banker, R., Potter, G., and Srinivasan, D. (2000), 'An Empirical Investigation of an Incentive Plan that Includes Nonfinancial Performance Measures', *The Accounting Review*, **75**(1), 65-92.
- Banker, R., and Datar, S. (2001), 'Sensitivity, Precision, and Linear Aggregation of Signals for Performance Evaluation', *Journal of Accounting Research*, **27**(1), 21-39.
- Barnow, Burt S. and Smith, Jeffrey A. (2002) "What Does the Evidence from Employment and Training Programs Reveal About the Likely Effects of Ticket-to-Work on Service Provider Behavior?" Working paper, June 18,.
- Berg, N., and Fast, N. (1975) 'Lincoln Electric Company', Harvard Business School Case 376-028.
- Burgess, Simon, Bronwyn Crosson, and Paul Gregg. (2001). The Intricacies of the Relationship Between Pay and Performance for Teachers: Do Teachers Respond to Performance Related Pay Schemes? Working Paper, CMPO, 01/035.
- Burgess, Simon, Carol Propper, and Debhorah Wilson. (2002). Does Performance Monitoring Work? A Review of the Evidence from the UK Public Sector, Excluding Health Care Working Paper, CMPO, 02/049.
- Burgess, Simon and Marisa Ratto. "The Role of Incentives in the Public Sector: Issues and Evidence. Oxford Review of Economic Policy. (This issue)
- Courty, P., and Marschke, G. (1997), 'Measuring Government Performance: Lessons from a Federal Job-Training Program', *American Economic Review*, **87**(2), 383-388.
- Courty, P. and Marschke, G. (2002). The Challenge of Measuring Productivity: A Case Study of Performance Measurement in a Job Training Program. Manuscript, University at Albany, State University of New York.
- Courty, P. and Marschke, G. (2003a). Performance Funding in Federal Agencies: A Case Study of a Federal Job Training Program. *Public Budgeting and Finance*. 1993, fall issues (Vol. 23:3).

Courty, P. and Marschke, G. (2003b). A Simple Test of Gaming. Manuscript, London Business School.

Courty, P., and Marschke, G. (2004, forthcoming), 'An Empirical Investigation of Gaming Responses to Explicit Performance Incentives', *Journal of Labor Economics*.

Darley, J. (1991), 'Setting Standards Seeks Control, Risks Distortion', *Institute of Government Studies Public Affairs Report*, **32**(4), Berkeley, University of California.

Dixit, A. (2002), 'Incentives and Organizations in the Public Sector', *Journal of Human Resources*, **37**(4), 696-727.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite. (2002) "Is More Information Better? The Effect of 'Report Cards' on Health Care Providers." NBER Working Paper 8697.

Feltham, G., and Xie, J. (1994), 'Performance Measure Congruity and Diversity in Multi-Task Principal/Agent Relations', *The Accounting Review*, **69**(3), 429-53.

Gibbons, R. (1997), 'Incentives and Careers in Organizations' in '*Advances in economics and econometrics: Theory and applications: Seventh World Congress*', (ed.), Kreps and Wallis, Cambridge University Press, 1997.

Goddard, M., Mannion, R. and Smith, P., (2000), "Enhancing performance in health care: a theoretical perspective on agency and the role of information", *Health Economics*, **9**, 95-107.

Goodhart, Charles. Central Banking, Monetary Theory and Practice : Essays in Honour of Charles Goodhart, Volume One. Edited by Paul Mizen. May 2003. Elgar, Edward Publishing.

Healy, P. (1985), 'The Effect of Bonus Schemes on Accounting Decisions', *Journal of Accounting and Economics*, **7**, 85-107.

Heckman, J. J., Smith, J., and Clements, N. (1997), 'Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts', *Review of Economic Studies*, **65**(4), 487-535.

Heckman, J. J., Heinrich, C., and Smith, J.A. (2002), 'The Performance of Performance Standards', *Journal of Human Resources*, **37**(4), 778-811.

Heinrich, Carolyn. (2002) "Outcome-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review*. **62**(6): 712-725.

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

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

Ittner, C. D., and Larcker, D. F. (1998), 'Are Nonfinancial Measures Leading Indicators of Financial Performance? An Analysis of Customer Satisfaction', *Journal of Accounting Research*, **36**, 1-35.

Jacob, Brian A., and Steven D. Levitt. 2002 Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. Manuscript, University of Chicago.

Kravchuk, R., and Schack, R. (1996), 'Designing Effective Performance-Measurement Systems under the Government Performance and Results Act of 1993', *Public Administration Review*, Jul/Aug, **56**(4), 348-358.

Loucks, W. and Whitney, W. (1973). *Comparative Economic Systems*, New York, NY: Harper & Row, ninth edition.

Marschke, G. (2001), 'The Economics of Performance Incentives in Government with Evidence from a Federal Job Training Program', in D. Forsythe (ed.), *Quicker, Better, Cheaper? Managing Performance in American Government*, Rockefeller Institute Press, pp. 61-97.

Oyer, P. (1998), 'Fiscal Year Ends and Non-Linear Incentive Contracts: The Effect on Business Seasonality', *Quarterly Journal of Economics*, **113**, 149-85.

Perrin, B. (1998), "Effective Use and Misuse of Performance Measurement," *American Journal of Evaluation*, **19**(3), 367-79.

Prendergast, C. (1999), 'The Provision of Incentives in Firms,' *The Journal of Economic Literature*, **37**(1), 7-63.

Prendergast, C. (2002), 'The Tenuous Trade-Off between Risk and Incentives', *Journal of Political Economy*, **110**(5), 1071-1102.

Smith, P. (1995) "On the unintended consequences of publishing performance data in the public sector", *International Journal of Public Administration*, **18**(2/3), 277-310.

Smith, Jeffrey. 1996. "A Note on Estimating the Relative Costs of Experimental and Non-Experimental Evaluations Using Cost Data from the National JTPA Study." Unpublished manuscript. The University of Chicago.

Propper, Carol and Debora Wilson. "The Use and Usefulness of Performance Measures in the Public Sector." *Oxford Review of Economic Policy*. (this issue)

Tirole, J. (1994), 'The Internal Organization of Government', *Oxford Economic Review*, **46**, 1-29.

Wholey, J., and Hatry, H. (1992), 'The Case for Performance Monitoring', *Public Administration Review*, **52**(6), 604-610.