PRIO
July 23, 2003
1:00 PM

Dynamics of Performance Measurement Systems

Pascal Courty
London Business School

Gerald Marschke
University at Albany, State University of New York [1]

First Draft: April 11, 2003

Preliminary and Incomplete
Do not quote

ABSTRACT: We present a model of how organizations manage performance measures when gaming is revealed over time. The incentive designer does not know when it selects a performance measure whether it will communicate the right behavior. Only over time does the principal find out the agent's responses and then uses this additional information to update and fine-tune the incentive system. Using data from a government organization, we test the model's main prediction that the correlation between a performance measure and the true goal of the organization should change after the performance measure is included in the incentive system and we find some evidence consistent with this hypothesis. (*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; and 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 for the support and encouragement of James Heckman.

1 Introduction

The economic literature on performance incentive design typically assumes that although organizations have to use imperfect performance measures, they know the relationship between these measures and the true organization's objective (Prendergast, 1999). In other words, the literature assumes that the principal knows how much gaming distortions will take place after the performance incentive system has been introduced. The principal uses this knowledge to trade-off the gains from incentives with the loss from distortions.

In practice, however, gaming is difficult to predict because it occurs after the agent acquires superior information about the technology of the performance measures. Some performance measures turn out to be easy to game while others are difficult. This suggests that the principal does not know how good a performance measure is until it is used. After a performance measure is used, however, unanticipated and unintended responses can be observed. If this is the case, then the principal should update the incentive system as more is learned about the effectiveness of different performance measures.

An important limitation of the incentive literature is that it provides only a static view of the design process of performance incentive systems. Accordingly, there is no need to modify the incentive system. This, however, this is inconsistent with the evidence on the functioning of performance incentive systems (Darley, 1991). There is 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 the Sears Auto Center which had to terminate its incentive system upon learning that it induced car mechanics to mislead consumers into unnecessary repair work. Incentive systems are not always 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).

This evidence suggests that there is a dynamic to incentive systems where principals do not know when they select a performance measure whether it will communicate the right behavior. Only over time does the principal find out the agent's responses and then uses this additional information to update and fine-tune the incentive system. Despite its empirical importance, this dynamic dimension is completely absent form the literature. The main goals of this paper are (a) to understand how organizations fine-tune incentive systems to control gaming costs and (b) to develop empirical tests of the dynamics of performance incentive systems.

We present a simple model of how organizations manage performance measures when gaming is revealed over time. The principal does not know ex-ante how much gaming a performance measure will generate but this is observed after the performance measure is used. An important insight from our analysis is that the statistical relation between a performance measure and the true goal of an organization is endogeneous. This relation is not the same before and after the performance measure is introduced in the incentive system. This implies that using a correlation measure to identify good performance measure can be misleading. One can discover how good a performance measure is only after it has been implemented and the gaming responses that it generates have been observed. This insight suggests that a selection method for performance measures that is based on how well measures predict the true objective (using correlation or other methods), as is commonly used by practitioners, has important limitations. In fact, such approach ignores that the statistical relation between the measure and the true goal does not account for the gaming actions that will take place after the measure is introduced.

Using data from a government organization, we test the model's main prediction that the correlation between a performance measure and the true goal of the organization should change after the performance measure is included in the incentive system. We find some evidence consistent with this hypothesis. An additional implication of the model is that poorly performing performance measures should be phased out. We find that the incentive designers seem to replace the performance measures that generate distortions.

This paper is organized as follows. The next section reviews the literature on the selection of performance measures. Section 3 presents a simple model and section 4 derives simple predictions. Section 5 tests some of the model's predictions in a government organization. Section 6 concludes.

2 Literature Review

This work builds on three strands of the literature on performance incentives. The first strand investigates the statistical relationship between the performance measures used in practice and the stated objective in the incentive systems. The second branch is the theoretical literature investigating how incentive systems should be designed. The third branch identifies dysfunctional responses to performance incentive. We show that there are gaps between these literatures and these gaps will be the focus of the rest of the paper.

2.1 Selection of Performance Measures

A sizeable portion of the management literature is devoted to understanding the design of incentive-based employment contracts, especially for private sector managers. An important pre-occupation of the literature is the selection of performance measures.[1] Much of the recent policy and public administration literature is also concerned with performance measurement, as interest in performance measurement and accountability has waxed in recent years (e.g. Wholey and Hartry (1992)).[2]

Researchers in these literatures test the validity of performance measures by correlating them with "true" measures of the goal of the organization. Measures that appear the most correlated with the goal are deemed most likely to be successful. To illustrate this point, we focus our review on the literature on the choice of performance measures in public sector job training because (a) this will be the topic of our empirical case study and (b) that literature illustrates nicely the points made elsewhere. Our review is based on the survey by Heckman, Heinrich, and Smith (2002). Their focus was on reviewing the evidence on whether short term performance measures are good predictors of long term impacts. We revisit their survey from a different perspective. We ask whether the relationship between short term performance outcomes and long term impacts depends on whether incentives are used at the time this relationship is measured.

[1] See Ittner and Larcker, 1998, for an example of a study that uses correlation methods to evaluate alternative performance measures for managerial compensation plans in the private sector.

[2] Burgess, Croxon, and Gregg (2001) present a review on the use of performance measurement in the UK public sector.

Job training programs that serve the economically disadvantaged have been an important part of the federal government's war on poverty at least since the Kennedy administration. The goal of such job training has been to raise the earnings ability of the economically disadvantaged. In the 1970s several influential studies showing the ineffectiveness of this job training forced Congress to reconsider how job training programs were constituted. In the early 1980s, some federal job training programs were restructured to allow job training administrators more discretion in the operation of their agencies. By allowing this discretion, Congress hoped that bureaucrats would be free to use their expertise in training and superior knowledge of "conditions on the ground" to provide better training. But in increasing bureaucrats' discretion over their work, Congress anticipated that administrators would also have greater means to pursue private objectives. Therefore, in addition to allowing more freedom in decision making, the program's overseers have sought to provide stronger incentives to promote job training's objectives by linking financial incentives to measures of program outcomes. These measures have been variants of program participants' employment and wage rates measured at the time they "graduated" from their training.

Numerous studies have attempted to test the validity of these employment-based measures by correlating them with earnings and employment gains at the individual enrollee level. These studies construct "true" measures of the success of job training, sometimes by exploiting data from social experiments run to assess the effectivness of job training, and sometimes by comparing the labor market success of persons who obtained training to persons from an artificially constructed control group. Some of these studies are run using data from job training programs that are subject to these performance-based measurement and others are not. Friedlander (1988) and Zornitsky, et al. (1988) conduct their studies based on data from job training programs that have no explicit performance measurement backed by financial incentives. They report that enrollees who are likely to produce high scores on employment-based performance measures are also likely to generate high earnings and employment impacts. Gay and Borus (1980), however, also using data from programs without incentive-backed performance measurement, found that the correlation of the employment measure and earnings impacts was sometimes negative.

Other studies based on data generated from job training programs subject to performance measurement such as Heckman, Heinrich, and Smith (2002) find evidence that the performance measures and earnings impacts are not significantly correlated and sometimes negatively correlated with gains. What is interesting for the purpose of our study, is that only studies of programs where performance is uncompensated show statistically significant correlation between performance measures and impacts.

2.2 Design of Incentive Systems

The theoretical literature on the design of incentives has developed around the principal-agent paradigm.[3] According to that paradigm, a principal hires an agent to complete a task but the agent has different preferences from the principal. The agent dislikes effort and the principal has to use performance incentives to influence the agent's choice of effort.

Holmstrom (1979) posits a risk neutral principal and a risk-averse agent and shows how the optimal contract allocates the output from the agent's effort between the principal and the agent. Because the agent's effort is unobservable, the principal and agent contract over a performance measure that is a noisy proxy of the agent's effort. Holmstrom shows that the optimal contract strikes a balance between insuring the agent from risk and providing incentives to elicit effort.

Multi-tasking models of the principal agent relationship (Holmstrom and Milgrom, 1991; Baker, 1992; Feltham and Xie, 1994; Banker and Datar, 1989) assume that aspects of the agent's value-added cannot be measured. As a consequence, in addition to balancing the provision of incentives and insurance, the optimal contract must also take into account the distortions caused by imperfect, mis-aligned measures of agent effort.

To illustrate this point consider the case of performance measurement in schools (Burgess, Propper and Wilson, 2002). In recent years some policy analysts and public officials have advocated setting up performance measures for local school districts, backed by federal educational subsidies as incentives. Such performance measures are

[3] The theoretical literature on the design of incentives is reviewed in Gibbons (1997) and Prendergast (1999). More recently, Dixit (2002) reviews the incentive literature but focusing on those issues that are specific to the public sector.

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 kinds of skills whose development is a principle aim of primary schools, however. Because they do not reward teaching citizenship, for example, theory predicts that teachers will neglect teaching citizenship. Instituting the performance measures can cause distortions by causing agents to spend no time on some activities that are productive but are not rewarded. Gaming—or taking actions which raise performance outcomes but that do not raise value-added—is another response to distortionary performance measures.

Implicit in these multitasking models is the idea that correlation between the performance measure and value-added is endogenous. Baker (2002) reviews the literature on distortion and risk and argues that the finding found elsewhere (Darley, 1991) that the correlation between performance measures and value-added degrade over time is consistent with multitasking models. He concludes that correlation is not a useful measure for selecting performance measures because it tells the incentive designer nothing about the gaming strategies available to the agent.

2.3 Dysfunctional Responses to Incentive

The empirical literature on gaming responses to incentives is surveyed in Gibbons (1997) and Prendergast (1999). A substantial fraction of this literature focuses on gaming response where the agent uses its discretion over the timing of performance reporting to meet performance thresholds. 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. More recent works report similar timing responses to threshold effects in other settings. For example, Asch (1990) showed that navy recruiters who receive awards for meeting year-end recruitment quotas respond by reallocating their work efforts over the year. Similarly, Oyer (1998) showed that there is more variability in firms' sales at the end of the fiscal years—when sales persons' bonuses are computed—than in the middle.

In the context of a training program, Courty and Marschke (2004) show that training program managers strategically time the reporting of their performance outcomes. An important distinction in their work is between unanticipated responses that diverts resources (e.g. agents' time) from productive activities and responses that simply reflects an accounting phenomenon. They find that the responses they identify are not simply an accounting phenomenon because they have a negative impact on the true goal of the organization.

There is also evidence of gaming in the schooling literature. Some scheme tie educational funding to scores on standarized tests and there is evidence that teachers have responded by "teaching to the test" and manipulating students' grade-to-grade promotions to boost scores. There have been a number of highly publicized incidents of teachers cheating (Jacob and Levitt, 2002).

2.4 Open Issues

There are two important gaps between the empirical and theoretical literatures on incentive design. First, the empirical literature on the selection of performance measures looks at correlations and does not control for the status of the performance measure in the incentive system at the time of measurement. This is in contradiction with the prediction of the theoretical literature that the statistical relation between the performance measure and value-added is endogeneous. According to the theoretical literature, the correlation between a performance measure and the true goal of the organization should change after the performance measure is introduced, a feature that has not been investigated in the empirical literature on the selection of performance measures.

Second, the theoretical literature has not incorportated important findings from the empirical literature on gaming. The literature on gaming suggests that it is likely that the principal does not know how much gaming a performance measure will generate until it is used. The theoretical literature has not incorporated this feature that the principal learns about gaming over time. Rather, the theoretical literature assumes that the principal knows how much gaming a measure will generate before it has been used. In that sense, the incentive literature assumes away the dynamic nature of the selection process of performance measures.

The goal of this paper is to fill these two gaps. First, we develop a simple model that studies how organization should design incentive systems when gaming is revealed over time. Second, we present some evidence on the changes in the statistical relation between performance measures and the true goal of the organization as new performance measures are introduced in the measurement system.

3 Model

The model builds on Baker's (1992) model of incentive design under imperfect performance measurement. We assume that the principal does not know how much gaming a performance measure will generate until the measure is used. We allow the principal to change the performance measures that generate too much gaming.

The principal hires an agent to invest in a project. The project is characterized by its type α, which is a random variable distributed with density f. The agent privately observes the project's type and invests in effort and in gaming. There are two imperfect performance measures. Both performance measures perfectly capture effort and both also have a gaming dimension

$$m_i(e, g_1, g_2; \alpha) = v(\alpha)e + w_i(\alpha)g_i$$

for $i = (1, 2)$. The first term captures investments that are perfectly aligned with the principal's true objective while the second term captures gaming distortions. Gaming investments are measure specific. The gaming actions that increase performance measure 1 leave performance measure 2 unchanged and vice versa. This is reasonable as long as the two performance measures are unlikely to share the same weaknesses. The costs of effort and gaming are the same across of all project types and are respectively $1/2e^2$ and $1/2g_i^2$, for $i = 1, 2$.

A performance measure can be of two types. A high gaming measure is such that $Ew^2(\alpha) = w^H$ while a low gaming one is such that $Ew^2(\alpha) = w^L$ with $w^H > w^L$. The principal observes whether a performance measure is a high or low gaming measure after it has been used. Before using a measure, the principal does not know its type but believes that it can be high or low with equal probability. Let $\bar{w} = 1/2(w^H + w^L)$.

There are two periods. In each period, the principal chooses one performance measure and a weight to put on it. We assume that it is not optimal to try out both performance measures in the first period. This is reasonable if there is a fixed cost of collecting data on a performance measure that is high enough so that it is never optimal to use both performance measures at the same time.

The principal uses linear contracts. Consider the incentive design problem with performance measure $m(e, g; \alpha) = v(\alpha)e + w(\alpha)g$. The agent's utility is,

$$U = E\left(\beta_0 + \beta m(e(\alpha), g(\alpha); \alpha) - 1/2e^2(\alpha) - 1/2g^2(\alpha)\right)$$

where the expectation is taken over density f and β is the weight on the performance measure and β_0 is a fixed payment. The agent's reservation utility is U_0. The principal's objective is the expected net return on the projects minus the expected payoff to the agent

$$\Pi = E\left(e(\alpha)v(\alpha) - (\beta_0 + \beta m(e(\alpha), g(\alpha); \alpha))\right).$$

Given the agent's investment response, the principal chooses the payment scheme (β_0, β) that maximizes Π subject to the constraint that the agent participates $U > U_0$.

We call $V(\alpha) = e(\alpha)v(\alpha)$ the realized objective and $M(\alpha) = e(\alpha)v(\alpha) + g(\alpha)w(\alpha)$ the realized performance outcome. The realizations of V and M depend on the incentive weight since this determines the agent's investment decisions.

4 Analysis

Before considering the two-period design problem, consider the optimal one-period contract with performance measure $m(e, g; \alpha) = v(\alpha)e + w(\alpha)g$. The agent's investment responses given contract (β_0, β) are given by the agent's first order conditions

$$e^*(\alpha, \beta) = \beta v(\alpha)$$

$$g^*(\alpha, \beta) = \beta w(\alpha).$$

The designer sets β_0 to satisfy 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

$$\Pi = E\left(\beta v^2(\alpha) - 1/2\beta^2 v^2(\alpha) - 1/2\beta^2 w^2(\alpha) - U_0\right).$$

The optimal incentive weight is

$$\beta = \frac{Ev^2}{Ev^2 + \bar{w}}.$$

When $Ew^2 = 0$, there is no gaming investment, the incentive weight equals one and the first-best is achieved. Otherwise, the optimal weight is lower than one because the principal anticipates that gaming will take place and accordingly reduces the weight on the performance measure. The performance measure can be expressed as a function of the principal's objective

$$M(\alpha) = V(\alpha) + \frac{Ev^2}{Ev^2 + \bar{w}} w^2(\alpha).$$

The performance measure is a sum of the principal's objective and a noise term that corresponds to gaming. The cost of gaming to the principal has two parts. First, there is a direct cost of gaming

$$1/2\left(\frac{Ev^2}{Ev^2 + \bar{w}}\right)^2 \bar{w}.$$

Second, there is an indirect cost of gaming. The threat of gaming implies that the principal lowers the incentive weight to trade-off efficient investment in effort and cost of gaming.

Next, we turn to the optimal two-periods contract. The principal tries out one measure in the first period, keeps that measure if it is a low gaming measure and switches to the other measure otherwise. Assume the principal tries performance measure 1 in the first period. The principal sets $\beta_2 = 0$. The optimal incentive weight on measure 1 is

$$\beta_1 = \frac{Ev^2}{Ev^2 + \bar{w}}$$

because the principal does not know if this measure is a high or low gaming measure.

In the second period, the principal observes the performance measure's type. If the performance measure is a low gaming measure, the principal increases the incentive weight

$$\beta_2 = \frac{Ev^2}{Ev^2 + w^L}.$$

If the performance measure is a high gaming measure, the principal switches measure at the end of period 1. In the second period, the weight on performance measure 1 is zero ($\beta_1 = 0$) and the weight on performance measure 2 is

$$\beta_2 = \frac{Ev^2}{Ev^2 + \bar{w}}.$$

This suggests the prediction that the weight on a performance measure should change over time in a systematic way.

Prediction 1 *The weight on a performance measure should either increase over time or the performance measure should be removed.*

Consider the case where the principal changes performance measure at the end of period 1. Let $M_i(\alpha, j)$ denote performance outcome i in period j. 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

$$M_2(\alpha; 1) = \frac{Ev^2}{Ev^2 + \bar{w}} v^2(\alpha).$$

But this is exactly equal to the principal's observed objective

$$V(\alpha; 1) = M_2(\alpha; 1).$$

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 more noisy since

$$M_2(\alpha; 2) = V(\alpha; 2) + \frac{Ev^2}{Ev^2 + \bar{w}} w_2^2(\alpha).$$

Prediction 2 *The true objective is a worse predictor of a performance measure when the performance measure is used as an incentive measure than when it isn't.*

The intuition is that gaming occurs only when the performance measure is included in the contract. When the principal moves from a high to a low gaming measure, the gain is equal to

$$1/2 \left(\frac{Ev^2}{Ev^2 + \bar{w}} \right)^2 (w^H - w^L).$$

5 Empirical Results

An important implication of our model is that the correlation between the performance measure and the goal of the organization is endogenous. That is, because placing incentives on performance measures cause agents to find low cost strategies to raise the performance measure that do not also raise the goal of the organization, the correlation between the performance measure and the goal of the organization degrades. In this section we outline a test of this response to the incentivization of performance measures in a real-world organization. The organization we study was created under the Job Training Partnership Act, until the late 1990s, one of the largest federal job training programs for the economically disadvantaged. It also was one of the first large-scale experiments with financially-backed performance incentives in a federal bureaucracy. In the mid-1980s, several years after the program began, the program's incentive designers changed the performance measures used to evaluate bureaucratic performance. We evaluate the relation between the goal of the organization and the new performance measures before and after their *exogenous* activation. The next subsection describes JTPA, its organization and the incentives in place.

5.1 JTPA

The Job Training Partnership Act of 1982 created what is was until 2000 the largest federal employment and training program serving the poor.[4] This section sketches the relevant features of JTPA.[5]

JTPA divided the U.S. into approximately 640 non-overlapping, collectively exhaustive jurisdictions. Within each sub-state region, a single administrative agency managed the region's budgetary appropriation. Agencies enjoyed wide discretion over who they enrolled, how many they enrolled, and the kinds of training they offered their enrollees. This has been noted elsewhere (see, for example, Courty and Marschke, 2002, and Marschke,

[4] In 2000, a new program created under the Workforce Investment Act (WIA) of 1998 supplanted JTPA. While many organizational details of WIA remain to be worked out, the change appears to be an evolutionary one. WIA retains the decentralized nature, the jurisdictional borders (see below), administrative entities, and the performance incentives of JTPA.

[5] See, *e.g.* Johnston (1987) for a detailed description of JTPA.

2002). To encourage the JTPA training centers to use their discretion to advance the goals of the act, the act required each state to pay out budgetary awards to successful training agencies from a fund equal to 6 percent of its annual appropriation. From year to year, the median training center won an award equal to about 7 percent of its budget, with some training centers winning nothing and others winning as much a sixty percent of their budget.

The act directed the U.S. Department of Labor (Department of Labor) to define performance measures to promote JTPA's mission. JTPA's stated mission was to increase the employment and earning ability of participants through job training (Section 106(a)).[6] Measuring worker output thus required knowing what each trainee would have earned in the absence of training, which of course is not observable. Good measures of counterfactual earnings and employment for everyday performance measurement use are prohibitively expensive to obtain. In the absence of cheap measures of human capital impact, the Department of Labor used performance measures based upon easily measurable labor market outcomes of trainees at or shortly after training.

These performance measures were based on aspects of the enrollee's labor market status at points in time. Early in the history of JTPA, labor market status was measured on the date the enrollee officially exited—or *terminated*—from the job training program. In the mid 1980s the incentive designers replaced measures based on labor market outcomes at the time of termination with measures based upon outcomes at 90 days after termination (see the discussion in section 5.4). Thus, JTPA eliminated the *employment rate at termination* and the *average wage at termination*, measured as the fraction of terminees over the course of the agency's fiscal year who were employed on their termination date and the average wage at the time of termination over those terminees who were employed at termination, respectively. In their place, the program's incentive designers installed similar measures based on employment status 90 days after termination. They replaced the employment at termination measure with two measures *employment rate at follow-up*

[6] Other goals mentioned in the Act include reducing welfare dependency among the poor (Section 106(a)), the equitable provision of services (Section
141(a)), and the placement of women in non-traditional jobs (Section 141(d)(2)). This study construes social value-added in terms of participant earnings and employment impacts only.

and *average weeks employed at follow-up*: measured as the employment rate at 90 days after termination and the number of weeks in the 90 days following termination averaged over the training agency's terminee population. They replaced the average wage at termination with *the average earnings at follow-up*, the total earnings in the 90 days following termination averaged over terminees who were employed at termination.

The incentive schemes were in effect for one year, which is coincident with the program's funding cycle, beginning on July 1 and ending on June 30 of the following calendar year. This unit of time is referred to as a *program year.* In any program year, all agencies within a state faced identical performance incentives. Between program years, states sometimes changed their performance measures—as they did in the mid-1980s at the DOL's direction. This exogenous variation in the performance measures from program year to program year is the basis of our empirical work.

5.2 Data and Method

The data source for this study is the National JTPA Study (NJS), an experimental study of the effectiveness of JTPA commissioned by the U.S. Department of Labor and conducted between 1987 and 1989. Sixteen of the organization's roughly 640 job agencies participated in the NJS.[7] The objective of the study was to produce estimates of the earnings and employment impacts of job training under JTPA that are free from selection bias. The study was conducted using a classical experiment methodology according to which persons who applied to the sixteen experimental training agencies between 1987 and 1989 were randomized into treatment and control groups. The control groups did not receive any JTPA training services for 18 months after random assignment. 20,601 JTPA-eligible adults and youth participated in the study: 13,972 were randomized into the treatment group and 6,629 into the control group.

The empirical analysis in this study is based on 13,338 usable adults from the set of participants in the NJS. The data contain participant-reported information on their education level, labor market history, family composition, welfare program participation

[7] See Doolittle and Traeger (1990) for a description of the implementation of the National JTPA Study, and Orr et al., 1994 for a detailed description of the its results.

and demographic characteristics, as well as labor market, training, and schooling activity for approximately 18 months after random assignment.[8] In addition, the data contain enrollment and termination dates for all experimental participants who also received training services. These program dates can be used with the participant employment spell, earnings and wage data to produce accurate measures of performance outcomes.

We follow closely the methodology of Heckman, Heinrich, and Smith (2002), who examine the correlation between JTPA's performance measures and the earnings and employment impacts of JTPA training using the same data we use here. We conduct separate analyses for each of three performance measures: the employment rate at follow-up, average weeks employed at follow-up, and average earnings at follow-up.

For each performance measure, we first divide the sample into two: one sample containing random assignments subject to the performance measure, the other sample containing random assignments not subject to the performance measure. To do this, we first identify for each training center in our data the program years for which the performance measure was in effect. The performance measures in place in each state and program year were obtained from documents on file in states' departments of labor. We then assign each experimental participant to the subsamples based on whether their random assignment date occurred in a program year in which their training agency was evaluated by the performance measure.

We cannot construct individual-specific earnings impacts using the experimental data (see Heckman, 1992, and Heckman, Smith, and Clements, 1997). Instead, following Heckman et al, we construct subgroup impacts. We construct 40 subgroups based on race, age, gender, family size, education, marital status, employment and earnings history, AFDC receipt, food stamp receipt, and welfare receipt. Thus, if an individual's data are complete he or she appears in our sample 40 times, but each individual appears in the data as many times as their data allow. For each individual in a subgroup, we compute an earnings figure by aggregating her earnings over the 18 months following their random

[8] For one quarter of the experimental participants, data were collected for an additional 18 months. This paper utilizes only the full sample, that is, only the employment data for the first 18 months following random assignment.

assignment. In the absence of a drop out problem, consistent estimates of the subgroup earnings impact can be obtained from a simple comparison of the 18 month earnings of treatments and controls within the subgroup. Over one-third of the individuals in the treatment group drop out, however. We use a regression framework to estimate the earnings impacts, employing a method suggested by Bloom (1984) to control for dropouts. We similarly compute employment impacts by comparing the number of months of employment reported by treatments and controls during the eighteen months following random assignment.

Table 1 shows the estimated earnings and employment impacts by subgroup for the sample of experimental participants upon which our analysis is based. Table 1 splits our sample into two subsamples, which differ by whether the experimental participants' random assignment dates correspond to a year in which their training agency was rewarded for high average weekly earnings at follow-up outcomes. Table 1 shows that the earnings impacts are in general small relative to their standard errors. This is consistent with findings based on these data reported elsewhere (e.g., Orr et al). Table 1 appears to show that impacts are higher after the performance measure is compensated. Compositional differences in the subsamples may also be responsible for the apparently higher earnings impacts when the average weekly earnings measure is compensated.

Because we compute earnings impacts by subgroup, we must compute performance measures by subgroup as well. Participants supplied monthly wage and employment information for each job held in the 18 month period after random assignment. The NJS data file also contains the exact enrollment and termination dates from agency records. Table 2 shows the average weekly earnings outcomes for selected subgroups by whether the performance measure is rewarded. We constructed the follow-up date-based performance outcomes using the enrollee's reported employment hours and wage information from the calendar month containing the termination date through the calendar month containing the follow-up date (the follow-up date occurs ninety days after the termination date). For the employment rate at follow-up measure, treatments were considered employed at follow-up if they showed employment in the third calendar month following termination. We constructed the average weekly earnings at follow-up measure by aggregating for each experimental enrollee the total labor market earnings they reported from the month of termination through the month of follow-up, inclusive. To be consistent with JTPA's definition of the measure, we constructed the earnings measure from only those persons who were employed in the third month following termination. To compute the subgroup performance outcomes, we averaged the individual performance outcomes within each subgroup. We constructed the average weeks worked similarly but instead of aggregating earnings over the follow-up period, we aggregated the number of weeks of employment over the follow-up period. Note that Table 2 shows that average weekly earnings at follow-up outcomes were in general higher after the measure was activated. This is consistent with our model of incentive response. Note also the substantial variation in the measure across subgroups.

5.3 Results

A primary implication of the model is that the correlation between the performance outcomes and the true goal of JTPA is endogenous. Following the literature, we use two measures for the true goal, earning impact and employment impact. We regress subgroup performance outcomes on their estimated employment and earnings impacts. We use regression, rather than correlations as suggested by the model, because our construction method for the subgroup observations implies that these observations are not independent (Heckman et al., 2002). In the spirit of the model, we test whether the coefficient on the impact falls with the activation of the measure. We take a finding that the coefficient falls as evidence that activating a performance measure weakens its association with programmatic impacts.

Table 3 shows the results of six regressions; there are two regressions for each of three performance measures, one regression for employment impact and one for earning impact. The dependent variable is either the subgroup outcome corresponding to the employment rate at follow-up, average weeks worked at follow-up, or average weekly earnings at follow-up performance measures. Each regression contains on the right hand side the estimated subgroup impact (either employment or earnings) and the impact interacted with dummy variable indicating whether the performance measure is activated. Each regression also contains an intercept and the activation dummy alone; these coefficient estimates are omitted from the table. Because the subgroups share some of the same persons, the regressions' observations are not independent. Therefore, the t statistics reported in Table 3 are computed with robust standard errors.

First, note that in five of the six regressions the coefficients on the estimated impacts are statistically indistinguishable from zero. That is, when they deactivated, the employment rate at follow-up and average weeks worked at follow-up outcomes are not correlated with either earnings or employment impacts. When deactivated, the average weekly earnings at follow-up outcome is only (marginally) correlated with the employment impact.

Second and more importantly, note that the activation of a performance measure lowers the association between the corresponding performance outcome and an impact in four of the six cases. In the case of the average weekly earnings at follow-up measure, activating the measure lowers the association between outcome and impact significantly.

6 Switching Performance Measures

A prediction of the model is that when the principal learns about gaming she will switch performance measures if she believes alternative measures would promote less gaming (Prediction 1). 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 give weight to efficiency considerations in the shaping of agency behaviour. Cost-based measures judged JTPA's managers by how much they spent to produce a job placement. Early in the program, local managers were rewarded for maintaining low expenditures per program participant. 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."[9] The cost measures' implementation and removal suggest an important dynamic in the construction of performance measures. JTPA's performance measure designers first instituted the cost measures to give some weight to efficiency considerations and 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. The JTPA designers introduced for the first time 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."[10] The switch from termination-date-based measures to follow-up measures constituted the second important change to the JTPA measurement system.[11]

Between 1992 and 2000, the year the Workforce Investment Act (WIA) supplanted JTPA, performance measures remained largely unchanged. With WIA, however, the performance measures were further refined, apparently taking into consideration, for ex-

[9] Federal Register January 5, 1990.

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

[11] There exist other changes, the circumstances around which 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 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. Under JTPA, Department of Labor officials closed this "loop hole" by limiting the time an idle enrolee could remain on the books to 90 days.

ample, the cream-skimming evidence developed during JTPA. As in JTPA, in WIA most of the performance measures are based on the labour market outcomes of enrolees. All labour market outcomes are measured after training ceases (as opposed to, on the date of termination), as in the latter day JTPA. However, WIA includes among the JTPA-style performance measures a new before-after measure of enrolees' earnings. Conceptually, the difference between an enrolee'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. Supposedly, before-after measures should lead to less cream skimming. The performance measures under WIA also include a measure of "customer satisfaction" produced from post-training surveys of enrolees and their employers.

7 Conclusion

This paper focuses on the selection of performance measures. We relax the assumption that is traditionally made in the incentive literature that the principal knows how much distortion a performance measure will generate before the performance measure is used. 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. An implication of our model is that selecting performance measures on the basis of their correlation with the organizations 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. We also show that the selection of performance measure follows a dynamic pattern where the principal has to try a performance measure to learn how much distortions it generates. Thus, incentive systems evolve over time where those performance measures that tend to generate too many distortions are phased out.

Using data from the JTPA incentive system, we test the model's main prediction that the correlation between a performance measure and the true goal of the organization should decrease after the performance measure is included in the incentive system. To test for existence of gaming, we focus on the introduction of the follow-up measures, which corresponds to one of the most dramatic changes in the measurement system. For three follow-up measures, we test whether the correlation between each measure and the true goal of the organization has decreased after the introduction of the measure. We find some evidence consistent with our hypothesis.

Our evidence suggests that using a correlation measure to identify good performance measures can be misleading. One can discover how good a performance measure is only after it has been implemented and the gaming responses that it generates have been observed. A selection method for performance measures that is based on how well measures predict the true objective (using correlation or other methods), as is commonly used by practitioners, has important limitations. A positive implication of our analysis is that one can discover how good a performance measure after it has been implemented and the gaming responses that it generates have been observed by testing for changes in correlation. This might be a valid approach to identify poor measures and to eliminate them.

References

[1] Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, 100(3):598–614.

[2] Baker, G. P. 2002. Distortion and Risk in Optimal Incentive Contracts. *Journal of Human Resources*, 37(4): 728-751.

[3] Baker, G. P., Robert Gibbons, and Kevin Murphy. 1994. Subjective Performance Measures in Optimal Incentive Contracts. *Quarterly Journal of Economics*, 109(2): 1125-1156.

[4] Banker, R. and Datar, S. 2001. Sensitivity, Precision, and Linear Aggregation of Signals for Performance Evaluation. *Journal of Accounting Research*, 27(1): 21-39.

[5] Bloom, H. S. (1984). Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review*, 8

[6] Burgess, Simon, Bronwyn Croxson, and Paul Gregg. (2001). The Intricates of the Relationship Between Pay and Performance for Teachers: Do Teachers Respond to Performance Related Pay Schemes? Working Paper, CMPO, 01/035.

[7] 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.

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

[9] Courty, P. and Marschke, G. (2002a). Implementing Performance Measurement and Performance Funding: Lessons from a Federal Job Training Program. Manuscript, University at Albany, State University of New York.

[10] Courty, P. and Marschke, G. (2002b). 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.

[11] Courty, P. and Marschke, G. (1997). Measuring Government Performance: Lessons from a Federal Job-Training Program. *American Economic Review*, 87(2):383–388.

[12] Darley, John. (1991). Setting Standards Seeks Control, Risks Distortion. *Institute of Government Studies Public Affairs Report*, 32(4). Berkeley: University of California.

[13] Dixit, Avinash. 2002. Incentives and Organizations in the Public Sector. *Journal of Human Resources*, 37(4): 696-727.

[14] Doolittle, F. and Traeger, L. (1990). *Implementing the National JTPA Study.* Manpower Demonstration Research Corporation, New York.

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

[16] Friedlander, D. 1988. *Subgroup Impacts and Performance Indicators for Selected Welfare Employment Programs.* New York: Manpower Development Research Corp

[17] Gay, R. and M. Borus. 1980. Validating Performance Indicators for Employment and Training Programs. *Journal of Human Resources*, 15, 1: 29-48.

[18] Gibbons, Robert. "Incentives and Careers in Organizations." In *Advances in economics and econometrics: Theory and applications: Seventh World Congress* University Press, 1997.

[19] Heckman, J. 1992. Randomization and Social Program Evaluation, in *Evaluating Welfare and Training Programs*, (C. Manski and I. Garfinkel ed.), 201-230. Cambridge, MA: Harvard University Press.

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

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

[22] 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.

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

[24] Johnston, J. W. (1987). *The Job Training Partnership Act: A Report by the National Commission for Employment Policy.*

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

[26] Marschke, G. (2002). Performance Incentives and Bureaucratic Behavior: Evidence from a Federal Bureaucracy. Manuscript, University at Albany.

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

[28] Orr, L. L., Bloom, H. S., Bell, S. H., Lin, W., Cave, G., and Doolittle, F. (March 1994). *The National JTPA Study: Impacts, Benefits, and Costs of Title II-A.* Abt Associates Inc.

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

[30] Wholey, Joseph and Harry Hartry. (1992). The Case for Performance Monitoring. *Public Administration Review*, 52 (6) 604-610.

[31] Zornitsky, J., Rubin M., Bell, S., and Martin, W. 1988 Establishing a Performance Management System for Targeted Welfare Programs. Washington DC: National Commission for Employment Policy, Research Report 88-14.

Table 1
Experimental Impacts
by Whether Average Weekly Earnings at Follow-up Measure Rewarded
for Selected Subgroups

	18 Months Earnings Impacts ($)		18 Months Employment Impacts (months)	
	AWEF Rewarded			
Subgroup	No	Yes	No	Yes
	Food Stamps			
Not Receiving Food Stamps	325.68	854.63	.1805	-.0517
	(381.48)	(607.79)	(.2156)	(.4092)
Receiving Food Stamps	365.01	994.23	.0586	1.1151
	(365.54)	(557.93)	(.2710)	(.4743)
	Gender			
Male	9.88	1065.40	.0727	.2761
	(444.89)	(906.01)	(.2393)	(.5416)
Female	560.05	852.97	.1103	.5967
	(301.82)	(426.73)	(.2354)	(.3767)
	Education			
Highest grade completed < 10 yrs	435.08	1443.40	.2448	1.0491
	(618.84)	(932.39)	(.4371)	(.6971)
Highest grade completed 10-11 yrs	564.47	1511.96	.0016	1.2830
	(528.15)	(807.19)	(.3749)	(.6299)
Highest grade completed 12 yrs	189.58	1137.65	.0410	.2919
	(410.46)	(672.08)	(.2691)	(.5068)
Highest grade completed 13-15 yrs	541.39	80.13	-.2016	-.9684
	(787.58)	(1773.04)	(.4301)	(1.0566)
Highest grade completed > 15 yrs	-1961.03	223.40	.9763	-.6607
	(2190.85)	(3016.34)	(.7840)	(1.7910)
	Race			
White	548.60	897.42	.0240	.2233
	(364.92)	(557.11)	(.2189)	(.3982)
Black	71.07	1125.39	.2248	.7492
	(493.13)	(676.14)	(.3297)	(.5405)
Hispanic	-697.12	4043.07	-.0682	2.9525
	(726.75)	(2447.82)	(.4801)	(1.4779)
Other	996.32	-1585.83	1.019	.3994
	(1394.74)	(2437.96)	(.9478)	(1.8558)

Notes: Robust standard errors of the estimates reported in parentheses. The estimated impacts are corrected for treatment group drop-outs. The earnings and employment impacts are estimated from the 10746 adult experimental participants who report a valid earnings figure (zeros are included) in each of the 18 months after random assignment. The employment impacts are denominated in months of employment and the earnings impacts are denominated in dollars.

Table 2
Average Weekly Earnings at Follow-up
by Whether Average Weekly Earnings at Follow-up Measure Rewarded
for Selected Subgroups

	AWEF Rewarded	
Subgroup	No	Yes
	Food Stamps	
Not Receiving Food Stamps	232.91	240.09
	(127.90)	(157.29)
Receiving Food Stamps	207.64	213.00
	(113.81)	(147.87)
	Gender	
Male	224.23	269.61
	(138.94)	(207.76)
Female	197.42	211.49
	(103.34)	(118.07)
	Education	
Highest grade completed < 10 yrs	206.99	216.98
	(115.23)	(132.86)
Highest grade completed 10-11 yrs	212.56	223.41
	(105.75)	(104.95)
Highest grade completed 12 yrs	221.68	233.98
	(118.42)	(196.98)
Highest grade completed 13-15 yrs	246.07	242.37
	(145.39)	(112.33)
Highest grade completed > 15 yrs	258.02	256.82
	(144.04)	(133.04)
	Race	
White	224.96	232.69
	(125.10)	(180.49)
Black	231.44	226.78
	(134.80)	(109.47)
Hispanic	200.84	245.31
	(98.58)	(104.08)
Other	242.72	218.86
	(118.16)	(76.43)
All	224.50	231.35
	(124.85)	(156.00)

Notes: Standard deviations are reported in parentheses. The number of observations used in each panel's calculation depends on the number of observations with valid responses. The calculations in the first (second) column are based on as many as 2465 (831) observations. See the text for an explanation of how average weekly earnings at follow-up is constructed.

Table 3
Outcome-Impact Regressions

	Dependent Variable		
Coefficient	Employment Rate at Follow-up	Average Weeks Worked at Follow-up	Average Weekly Earnings at Follow-up
Earnings Impact	.000005	.000011	.007
	(.166)	(.046)	(.574)
Earnings Impact	-.000015	-.000068	-.0218
X Activation Dummy	(-.480)	(-.261)	(-1.638)
R^2	.006	.079	.340
Employment Impact	-.020	-.179	20.042
	(-.785)	(-.477)	(1.935)
Employment Impact	.012	.172	-42.160
X Activation Dummy	(.367)	(.421)	(-3.261)
R^2	.013	.080	.410

Notes: T statistics in parentheses based on robust standard errors. Activation dummy coded as one if the relevant performance measure in effect, as zero otherwise. The constant and coefficient on the activation dummy are omitted.