Dynamics of Performance Measurement Systems

Pascal Courty
London Business School

Gerald Marschke
University at Albany, State University of New York

First Draft: November 7, 2002

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Do not quote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.

1Comments 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 from 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. Two additional implications of the model are that (a) performance measures are selected on the basis of how well they communicate the organization’s goal and (b) poorly performing performance measures should be phased out. We find mixed evidence in support of these hypothesis. On the one hand, performance measures are not selected on the basis of their correlation with the true goal of the organization. On the other hand, the incentive designers seem to replace the performance measures that generate distortions. One way to interpret this finding is
that the incentive designers in our case study may have other concerns in addition to the goal stated in the organization’s mission.

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

¹See 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.
²Burgess, Croxon, and Gregg (2001) present a review on the use of performance measurement in the UK public sector.
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.

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 effectiveness 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

---

³The 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.
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 pub-
lic officials have advocated setting up performance measures for local school districts,
backed by federal 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 kinds of skills whose development is a principle aim of primary schools, how-
ever. 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 perfor-
mance 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 consis-
tent 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 re-
response where the agent uses its discretion over the timing of performance reporting to
meet performance thresholds. Healy (1985) documents that managers who are compen-
sated for meeting annual income thresholds use their discretion over the timing of income
reporting to smooth their compensation across accounting years. More recent works re-
port 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 incorporated 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 $\alpha$, 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 \( \mathbb{E}w^2(\alpha) = w^H \) while a low gaming one is such that \( \mathbb{E}w^2(\alpha) = w^L \) 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,

\[
U = \mathbb{E}(\beta_0 + \beta m(e(\alpha), g(\alpha); \alpha) - 1/2e^2(\alpha) - 1/2g^2(\alpha))
\]

where the expectation is taken over density \( f \) and \( \beta \) is the weight on the performance measure and \( \beta_0 \) 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

\[
\Pi = \mathbb{E}(e(\alpha)v(\alpha) - (\beta_0 + \beta m(e(\alpha), g(\alpha); \alpha))).
\]

Given the agent’s investment response, the principal chooses the payment scheme \((\beta_0, \beta)\) that maximizes \( \Pi \) 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 \((\beta_0, \beta)\) are given by the agent’s first order conditions

\[ e^*(\alpha, \beta) = \beta v(\alpha) \]
\[ g^*(\alpha, \beta) = \beta w(\alpha). \]

The designer sets \(\beta_0\) to satisfy the agent’s participation constraint. Replacing \(\beta_0\) 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

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

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. 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

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

**Prediction 1** 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
\[ \frac{1}{2} \left( \frac{E v^2}{E v^2 + w} \right)^2 (w^H - w^L). \]

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.\(^5\)

\(^4\)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.
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, 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)). 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

---

6Other 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.
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 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

---

7See 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.
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 and demographic characteristics, as well as labor market, training, and schooling activity for approximately 18 months after random assignment. 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. 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

\footnote{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.}

\footnote{The performance measures in place in each state and program year were obtained from documents on file in states’ departments of labor.}
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 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 drop-outs. 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 not 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 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 from 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 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.

5.3 Results

A primary implication of the model is that the correlation between the performance outcome and earnings and employment impact is endogenous. In the spirit of the model, we test whether the correlation between the performance measure and the program impact varies by whether the performance measure is compensated or not.

Table 3 shows the simple correlations between the impact estimates and performance outcomes before and after their activation for the employment rate at follow-up, average weeks worked at follow-up, and average weekly earnings at follow-up performance measures. For example, the first line in the first panel of the table shows the Pearson correlation coefficients for estimated earnings impacts and the employment rate at follow-up performance outcome. The first and second columns show the the correlations before and after the performance measure is rewarded, respectively. We report the p value of the tests of the correlations’ significance in parentheses below their correlations. The third column of the table reports the p value under the null that the correlations are equivalent. The third line in the table repeats the analysis of the employment rate at follow-up measure for the employment impact measure. The other two panels of the table display the results of the analysis of the average weeks worked at follow-up and average weekly

\[^{10}\text{To be consistent with JTPA’s definition of the measure, we constructed the earnings measure from only those persons who were employed in the third month following termination.}\]
earnings at follow-up measures.

First, note that in no case is the correlation between the performance outcome and the estimated impact positive and significant before the performance measure is used.\textsuperscript{11} If one takes our theoretical model seriously, this presents a puzzle, as the performance measure in the absence of gaming must be equal to value-added. In our opinion the likely explanation is that JTPA’s incentive designers did not initially choose performance measures on the basis of their correlation with value-added, but to meet some other objective.

Second, note that in most cases the correlation between the performance outcome and program value-added falls with the implementation of the performance measures. Consider first earnings impacts. The table shows that the correlation with each performance measure is positive before the measure’s activation but negative after it. In the case of the average weekly earnings at follow-up measure, the correlation after is significantly negative and the difference between the before and after correlations is also significant. In the case of employment impacts, while in two of three cases the correlation moves from positive to negative, in no case is the change in correlation significant.

The flip in the signs of the correlation coefficient is difficult to reconcile with our model. Our model implies that the performance measure is equal to value-added before the measure is rewarded. After the measure is rewarded, however, the performance measure is equal to value added plus a positive term that is proportional to the agent’s gaming activity. This introduces new variation in the performance measure that is orthogonal to the variation in value-added, causing the correlation between value-added and the performance outcome to fall \textit{toward zero}, but not become negative.

A different model than the one typically used in the gaming literature, and exposited in this paper, could explain our result. The literature has typically assumed that gaming activities and non-gaming activities entered in an additive fashion in the production technology of the performance measure. An alternative hypothesis would be that gaming and non-gaming are exclusive in the sense that gaming precludes investment in other activities. Suppose the training agency faces a trade-off between gaming and human-capital

\textsuperscript{11}When the average weekly earnings at follow-up is not rewarded, its correlation with employment impacts are marginally significant by conventional significance criteria.

building effort, so that effort devoted to gaming came necessarily at the expense of human-capital building. Before the program’s designers implemented the performance measure we would observe perfect correlation between $M$ and $V$. Now, after the performance measure was implemented, the training agency would begin devoting some effort to gaming. Subgroups that received higher gaming would also show higher performance outcomes, and because gaming is destructive, lower earnings gains. Thus program impacts and performance become negatively correlated.

### 5.4 Switching Performance Measures

An implication of our model is that when the principal learns about gaming she will switch performance measures if she believes alternative measures would promote less gaming. As discussed in section 5.1, the program’s original measures were termination-based. The Department of Labor’s phase out of termination-based performance measures in the mid-1980s was in response to investigations by the General Accounting Office that seemed to show that these measures, with their emphasis on the enrollee’s employment state on the last day of training, induced training agencies to emphasize “quick fixes”, that is, job placement-oriented services that had no long-term impact on enrollees’ skills. By introducing measures based on the employment state of enrollees three months after the official end of their association with the training centers—or follow-up measures—to “[promote] effective service to participants and [assist] them to achieve long-term economic independence.” The switch from termination-date-based measures to follow-up measures constituted one of the main changes to the JTPA measurement system and appears to have been precipitated by the suspicion that termination-based measures were being gamed.

Findings in Courty and Marschke, 2002b, may explain our results in Table 3. They present informal evidence that training agencies responded to the switch to follow-up mea-

---

12The incentive system in the Sears Auto Center example probably exhibited this trade off between the gaming activities and the objective of the firm. In the Sears example, gaming activities—selling customers unnecessary repairs—increased short-run sales-based performance measures, but they reduced the value of the firm because they damaged its reputation and lowered future sales.

13The discussion in this subsection is based on Courty and Marschke, 2002b.

sures by offering unemployed terminees job counseling and placement services throughout the follow-up period. Courty and Marschke argue that these services were themselves quick-fixes, designed to inflate the new performance outcomes and producing no lasting positive effect on the enrollee. Courty and Marschke note that these offers of job counseling and placement ceased at the end of the follow-up period. Courty and Marschke thus offer evidence that strategies were available for follow-up measures that inflated the performance outcome without affecting enrollees’ human capital levels and that these strategies were implemented only after states rewarded the measure.

6 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. Relaxing this assumption has important implications for the design of incentive systems. We study how principal select performance measures and following the empirical literature we ask whether a selection rule that focuses on correlation is valid. We show that this rule may be flawed because it ignores gaming. 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 a government organization, we test and find evidence consistent with 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. Another implication of the model is that performance measures are selected on the basis of how well they communicate the organization’s goal. We find mixed evidence in support of this hypothesis. One way to interpret this finding is that the incentive designers in our case study may have other concerns in addition to the goal stated in the organization’s mission.
References


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

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>18 Months Earnings Impacts ($$)</th>
<th>18 Months Employment Impacts (months)</th>
<th>AWEF Rewarded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Food Stamps</td>
<td>Food Stamps</td>
<td></td>
</tr>
<tr>
<td>Not Receiving Food Stamps</td>
<td>325.68</td>
<td>854.63</td>
<td>.1805</td>
</tr>
<tr>
<td></td>
<td>(381.48)</td>
<td>(607.79)</td>
<td>(.2156)</td>
</tr>
<tr>
<td>Receiving Food Stamps</td>
<td>365.01</td>
<td>994.23</td>
<td>.0586</td>
</tr>
<tr>
<td></td>
<td>(365.54)</td>
<td>(557.93)</td>
<td>(.2156)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>9.88</td>
<td>1065.40</td>
<td>.0727</td>
</tr>
<tr>
<td></td>
<td>(444.89)</td>
<td>(906.01)</td>
<td>(.2393)</td>
</tr>
<tr>
<td>Female</td>
<td>560.05</td>
<td>852.97</td>
<td>.1103</td>
</tr>
<tr>
<td></td>
<td>(301.82)</td>
<td>(426.73)</td>
<td>(.2354)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highest grade completed &lt; 10 yrs</td>
<td>435.08</td>
<td>1443.40</td>
<td>.2448</td>
</tr>
<tr>
<td></td>
<td>(618.84)</td>
<td>(932.39)</td>
<td>(.4371)</td>
</tr>
<tr>
<td>Highest grade completed 10-11 yrs</td>
<td>564.47</td>
<td>1511.96</td>
<td>.0016</td>
</tr>
<tr>
<td></td>
<td>(528.15)</td>
<td>(807.19)</td>
<td>(.3749)</td>
</tr>
<tr>
<td>Highest grade completed 12 yrs</td>
<td>189.58</td>
<td>1137.65</td>
<td>.0410</td>
</tr>
<tr>
<td></td>
<td>(410.46)</td>
<td>(672.08)</td>
<td>(.2691)</td>
</tr>
<tr>
<td>Highest grade completed 13-15 yrs</td>
<td>541.39</td>
<td>80.13</td>
<td>-.2016</td>
</tr>
<tr>
<td></td>
<td>(787.58)</td>
<td>(1773.04)</td>
<td>(.4301)</td>
</tr>
<tr>
<td>Highest grade completed &gt; 15 yrs</td>
<td>-1961.03</td>
<td>223.40</td>
<td>.9763</td>
</tr>
<tr>
<td></td>
<td>(2190.85)</td>
<td>(3016.34)</td>
<td>(.7840)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>548.60</td>
<td>897.42</td>
<td>.0240</td>
</tr>
<tr>
<td></td>
<td>(364.92)</td>
<td>(557.11)</td>
<td>(.2189)</td>
</tr>
<tr>
<td>Black</td>
<td>71.07</td>
<td>1125.39</td>
<td>.2248</td>
</tr>
<tr>
<td></td>
<td>(493.13)</td>
<td>(676.14)</td>
<td>(.3297)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-697.12</td>
<td>4043.07</td>
<td>-.0682</td>
</tr>
<tr>
<td></td>
<td>(726.75)</td>
<td>(2447.82)</td>
<td>(.4801)</td>
</tr>
<tr>
<td>Other</td>
<td>996.32</td>
<td>-1585.83</td>
<td>1.019</td>
</tr>
<tr>
<td></td>
<td>(1394.74)</td>
<td>(2437.96)</td>
<td>(.9478)</td>
</tr>
</tbody>
</table>

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  

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

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>
<thead>
<tr>
<th>Measure Rewarded</th>
<th>Employment Rate at Follow-up</th>
<th>p value from test of equality of correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Employment Impact</td>
<td></td>
</tr>
<tr>
<td>Earnings Impact</td>
<td>.0041</td>
<td>-.1024</td>
</tr>
<tr>
<td></td>
<td>(.980)</td>
<td>(.529)</td>
</tr>
<tr>
<td>Employment Impact</td>
<td>-.1405</td>
<td>-.0688</td>
</tr>
<tr>
<td></td>
<td>(.387)</td>
<td>(.673)</td>
</tr>
<tr>
<td></td>
<td>Average Weeks Worked at Follow-up</td>
<td></td>
</tr>
<tr>
<td>Earnings Impact</td>
<td>.0086</td>
<td>-.0722</td>
</tr>
<tr>
<td></td>
<td>(.958)</td>
<td>(.658)</td>
</tr>
<tr>
<td>Employment Impact</td>
<td>-.0933</td>
<td>-.0066</td>
</tr>
<tr>
<td></td>
<td>(.566)</td>
<td>(.968)</td>
</tr>
<tr>
<td></td>
<td>Average Weekly Earnings at Follow-up</td>
<td></td>
</tr>
<tr>
<td>Earnings Impact</td>
<td>.1604</td>
<td>-.6765</td>
</tr>
<tr>
<td></td>
<td>(.323)</td>
<td>(&lt;.001)</td>
</tr>
<tr>
<td>Employment Impact</td>
<td>.2843</td>
<td>-.7313</td>
</tr>
<tr>
<td></td>
<td>(.075)</td>
<td>(&lt;.001)</td>
</tr>
</tbody>
</table>

The p value under null that correlation coefficient is zero reported in parentheses.
The unit of observation is the subgroup. Each correlation is based on 40 subgroups.