A General Test of Gaming

Pascal Courty and Gerald Marschke[1]

June 2004

Abstract: An important lesson from the incentive literature is that explicit incentives may elicit dysfunctional and unintended responses, also known as gaming responses. The existence of these responses, however, is difficult to demonstrate in practice because this behavior is typically hidden from the researcher. We present a simple model showing that one can identify gaming by estimating the correlation between a performance measure and the true goal of the organization before and after the measure has been activated. Our hypothesis is that gaming takes place if this correlation decreases with activation. Using data from a public sector organization, we find evidence consistent with our hypothesis. We draw implications for the selection of performance measures.

JEL H72, J33, L14

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

[1] Pascal Courty, Department of Economics, European University Institute, Villa San Paolo, Via Della Piazzuola 43, 50133, Firenze, Italy. Email: Pascal.Courty@IUE.it; 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 thank seminar participants at the Tinbergen Institute, Turin, and at the 2003 NBER Summer Institute Productivity/Organizational Economics/Industrial Organization Joint Meeting, and Robert Gibbons, James Heckman, Carolyn Heinrich, and Jeff Smith for insightful comments. The usual disclaimer applies.

1 Introduction

Explicit performance measures may elicit dysfunctional and unintended responses, also known as gaming responses (Holmstrom and Milgrom 1991 and Baker 1992), which often prove costly to organizations. For examples of gaming responses and the damage they cause in the private sector, see the specific references to performance pay and employee misconduct (at Sears', among other firms) in Baker et al (1994). For an example from the public sector, see the study of teacher cheating in high schools by Jacob and Levitt (2002). Understanding when gaming responses take place, the extent of these responses and their nature, is essential to rule out poor incentive designs that could put the organization at risk and also more generally to improve the effectiveness of measurement systems. Gaming behavior, however, is difficult to identify because it is typically hidden from the researcher and in many cases (at least for some time) from the organization as well.

Despite this difficulty, a growing literature has demonstrated the existence of gaming in several organizational contexts. See Prendergast (1999) for a review. Following the seminal work of Healy (1985), this literature has circumvented the identification challenges by focusing on responses where gaming can be unambiguously identified from the specifics of the contract (e.g. manipulating accounting figures). The main shortcoming of this approach is that by its very case-study nature, it can be applied only to a narrow set of gaming responses and it requires detailed information on the contracts and on agent behavior that is often difficult to observe. There is no general method to address the question of whether a performance measure generates gaming.

We develop a new approach to identify gaming. Our starting assumption is that different performance measures generate different gaming responses. Although performance measures share in common the feature that they attempt to communicate the true organizational goal, they are imperfect proxies and their source of imperfections is likely to differ. The investment strategies that optimally game a given measure may have little impact elsewhere. We propose to estimate gaming strategies by assessing changes in performance outcomes before and after the introduction of a performance measure.

We extend Baker's 2002 gaming model to derive a simple test of gaming that only requires estimating how the correlation between a performance measure and the true goal of the organization changes after the measure has been activated in the incentive system. We show that gaming takes place if this correlation decreases after the introduction of the new performance measure. The intuition for this test is that after a measure is activated, the agent takes measure-specific actions that maximize the measure but that do not maximize the true goal. These actions increase the variability of the measure thus reducing the correlation between the measure and the true goal.

We test this hypothesis in the incentive system of a federal training organization created under the Job Training Partnership Act (JTPA), which operated from 1982 to 2000. There are several reasons for choosing this case study. To start, JTPA used incentive-backed performance measurement, assessing job training output with measures of the labor market success of training participants and rewarding successful managers with small budgetary increases. In addition, JTPA was the object of a large-scale experimental study that produced unusually precise and complete information on performance outcomes and organizational value. Another important reason for using this case study is that in the late 1980's the Department of Labor introduced new measures. For these new measures, we can observe performance outcomes as well as organizational outcomes, before and after the measure's introduction.

JTPA's stated objective was to raise the earnings ability and lower the welfare dependency of the poor. JTPA evaluated local managers' performance by their clients' labor market success (e.g. employment status) at the end, or shortly after the end of training. We test whether the correlation between the new performance measures and the true goal of the organization decreased after the introduction of these measures. We find that in all new performance measures a decline in correlation as predicted by the model.

This paper contributes to two literatures. First, and as mentioned earlier, it contributes to the literature trying to demonstrate the existence of gaming responses. A substantial fraction of the literature focuses on gaming responses where the agent uses its discretion over the timing and reporting of performance outcomes to meet performance thresholds (Healy, 1985; Asch, 1990; Oyer, 1998; Jacob and Levitt, 2002; Oettinger, 2002; Burgess et al., 2002; and Courty and Marschke, 2004). In contrast, our approach offers a more general test of gaming that only requires computing correlations before and after the introduction of a new measure. The main advantages of our approach are that it is general and it relies on data that can be easily collected.

This paper also contributes to the literature on the implementation of performance measurement. An important pre-occupation of the literature is the selection of performance measures (e.g., Gibbs et al, 2004). Researchers in this literature, and practitioners as well, evaluate the usefulness of performance measures based on how correlated they are with the true objective of the organization. Ittner and Larcker (1998), Banker, Potter, and Srinivasan (2000), and van Praag and Cools (2001), for example, use correlation methods to evaluate alternative performance measures for managerial compensation plans in the private sector. Much of the recent policy and public administration literature is also concerned with performance measurement, as interest in performance measurement and accountability in the public sector has waxed in recent years (e.g. Heckman, Heinrich, and Smith, 2002). 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. These methods, however, lack a theoretical justification. By showing that a validation method based on correlation prior to the measure' s introduction is flawed because it fails to capture the gaming strategies available to the agent, our model is a contribution to the practice of performance measurement validation and selection. Although one has to be cautious in using a correlation criterion to select performance measures, one can use the change in correlation to identify whether gaming takes place after a measure is introduced.

This paper is organized as follows. Section 2 presents a simple framework to test for gaming. Section 3 presents some preliminary evidence consistent with this framework and tests our predictions in the JTPA organization. Section 4 concludes.

2 Model

We adopt the multi-tasking principal agent paradigm (Holmstrom and Milgrom, 1991; Baker, 2002; Feltham and Xie, 1994; Banker and Datar, 2001) building upon our previous work (Courty and Marschke, 2003a).[2] In contrast to these works, which focus on the principal's problem of choosing the optimal contract, we focus on the agent's decision problem because the goal of the analysis is to investigate how the agent responds differently to different sets of performance measures.

To keep matter simple, we assume that there are two performance measures. We investigate whether the agent's responses to performance measure one differ in a systematic way from her responses when performance measure two is introduced, and whether one can identify the existence of gaming from the differences in responses. We assume that the introduction of performance measure two is exogenously given. A possible interpretation is that the principal does not know ex-ante which measure is likely to perform well. Different principals experiment with different measures and this generates exogenous variations in performance measures.

A different interpretation, which matches our application, is that the principal first uses performance measure one and later introduces performance measure two. Under that interpretation, the introduction of measure two could be endogenous. That is, it could be triggered by the

[2] 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.

observation of the agent's responses to measure one. Then, the agent may internalize the fact that her early actions may trigger a change in performance measures. Whether the agent makes such anticipations depends on the application. In the model we present, we ignore this possibility because the presence of more than independent 600 agents in our application implies that each agent has very little influence on the principal's decision to introduce a new measure.[3]

The agent invests in tasks. One could think of a task as a project. In the context of JTPA, for example, a task could be a single enrollee or a group of enrollees. The agent has to allocate resources across enrollees and the issue is how different performance measures change the agent's resource allocation. Each task is characterized by its type α. The agent privately observes the task's type α and invests in effort and gaming. By assumption, the principal values only effort.

For each task, the agent chooses a vector of investments $(e,g)=(e_0,e_1,e_2,g_1,g_2)$. The performance outcome for task α on measure i=1,2 is

$$p_{i,\alpha}(e,g)=v_{0,\alpha}e_0+v_{i,\alpha}e_i+w_{i,\alpha}g_i.$$

Our specification ignores performance measurement noise. This assumption is not restrictive for the analysis, which focuses on gaming rather than on the optimal weighting of performance measures.[4]

The principal's objective or social value-added of investment vector (e,g) on task α is,

$$V_\alpha(e,g)=v_{0,\alpha}e_0+ v_{1,\alpha}e_1+ v_{2,\alpha}e_2.$$

This specification recognizes the distinction made in the literature between multi-tasking and gaming.[5] The common dimension of effort e_0 captures what both performance measures and the

[3] In a single agent model, the agent anticipates the impact of her actions in stage one on the probability that the performance measure could be changed in stage two. The equilibrium dynamic of gaming investment and the decision to change the measure would be more complex. The main predictions of the model would likely follow, however, as long as in equilibrium the principal sometimes changes the set of performance measure used, which is the departing point of this research.

[4] In the standard principal agent model, measurement noise plays a role in the determination of the optimal contract, but it does not directly influence the agent's investment decisions.

[5] Our specification is very similar to the specification in Baker (2002), who assumes that, $V=f.a+\varepsilon$ and $P=g.a+\phi$, where f and g are vectors of marginal products of actions, a, in the principal's objective and performance outcome equations. We ignore the error terms ε and ϕ as they do not influence the agent's action choice (see previous footnote). In

true goal share in common. In addition to that, both performance measures imperfectly capture some dimensions of effort (multi-tasking) and both also have a gaming dimension. Multi-tasking is captured by the fact that each measure captures only one of the two effort margins e_1 and e_2. Gaming is captured by the margins g_1 and g_2 which increase the performance measures but not the true goal.[6] Gaming investments are measure specific. The gaming actions that increase performance measure one leave performance measure two unchanged and vice versa. This is reasonable as long as the two performance measures are unlikely to share the same weaknesses. Finally, to draw empirical predictions, we will assume the tasks are randomly drawn from the α population. To simplify, we will assume that the $v_{i,\alpha}$ and $w_{i,\alpha}$ are orthogonal to one another.

The focus of this paper is on identifying the existence of gaming responses g_1 and g_2. We say that performance measure i is gameable if $w_{i,\alpha}>0$ for some α.

The costs of effort and gaming are the same across all tasks and are respectively $½e_i^2$ and $½g_i^2$, for i=1,2. In our model the performance outcome for measure i is the sum of performance outcomes over all tasks

$$P_i=\sum_\alpha p_{i,\alpha}(e_\alpha,g_\alpha).$$

Assume for now that the weights on performance measure one and two are β_1 and β_2, respectively. The agent chooses effort and gaming investment (e_α,g_α) to maximize

$$\sum_i\beta_i[M_i-½\sum_\alpha(e_{0,\alpha}^2+\sum_i(e_{i,\alpha}^2+g_{i,\alpha}^2))].$$

To simplify the exposition, let $e_{i,\alpha}(\beta_1,\beta_2)$ and $g_{i,\alpha}(\beta_1,\beta_2))$ denote the optimal investment strategy for task α when the performance weights are β_1 and β_2 respectively and similarly denote $p_{i,\alpha}(\beta_1,\beta_2)$ and $V_\alpha(,\beta_1,\beta_2)$ the performance outcomes and principal's objective. The agent's investment response is

addition, we assume that each element of the vector of action can be decomposed into effort, multi-tasking, and gaming components $a=(e_0,e_1,e_2,g_1,g_2)$. By linearity, this is equivalent to assuming that one can separate the tasks that enter (f,g) into three components: perfectly aligned, multi-tasking and gaming.

[6] One could assume that some gaming activities have a destructive effect on the principal's objective. Formally, g would enter negatively in $V_\alpha(e,g)$. This would reinforce our main prediction but complicate the exposition without adding any new insights.

$$e_{0,\alpha}(\beta_1,\beta_2)= (\beta_1+\beta_2)v_{0,\alpha},$$

$$e_{i,\alpha}(\beta_1,\beta_2)= \beta_i v_{i,\alpha}, \text{ for } i=1,2$$

$$g_{i,\alpha}(\beta_1,\beta_2)= \beta_i w_{i,} \text{ for } i=1,2.$$

Investment $(e_\alpha(\beta_1,\beta_2),g_\alpha(\beta_1,\beta_2))$ generates the realized performance outcome for measure i

$$p_{i,\alpha}(\beta_1,\beta_2)=e_{0,\alpha}(\beta_1,\beta_2)v_{0,\alpha}+e_{i,\alpha}(\beta_1,\beta_2)v_{i,\alpha}+g_{i,\alpha}(\beta_1,\beta_2)w_{i,\alpha}. \quad (1)$$

The principal's realized objective for task α is,

$$V_\alpha(\beta_1,\beta_2)=e_{0,\alpha}(\beta_1,\beta_2)v_{0,\alpha}+\sum_i e_{i,\alpha}(\beta_1,\beta_2)v_{i,\alpha}.$$

The realized performance outcome and the realized objective depend on the agent's investment, which in turn depends on which measure is activated and on the performance weights.

The model makes several simplifying assumptions. The central assumption of the model is that different measures are likely to display different gaming weaknesses. As we will see, conditional on this assumption it is possible to identify measure-specific gaming responses. In addition, the model assumes that (a) the performance measures and organizational goal are linear in the agent's actions and (b) the marginal products of actions are independent. Although simplistic, these assumptions are made for tractability and should be interpreted as a first order approximation of a more complex specification.

Performance Measure Activation

Organizations often change the performance measures they use, sometimes replacing outdated ones, or augmenting the old measures with new ones as they become available. Here we consider the latter case. Extending the analysis to the former case, however, where one performance measure is replaced by another, yields similar implications.

Assume that the principal first activates only performance measure one and then decides to activate measure two. The performance weights change from $(\beta_1,0)$ to (β_1',β_2'). Do performance outcomes change in a systematic way as the set of activated measure changes? Expression (1) suggests a prediction on how the mean of a performance measure should change after it is activated. If overall incentive weights do not decrease, $\beta_1' + \beta_2' \geq \beta_1$, the mean performance outcome on measure two should increase when that measure is activated. Similarly, the variance of measure two should increase after its introduction,

$$\mathrm{Var}\ p_{2,\alpha}(\beta_1',\beta_2') = \mathrm{Var}(\beta_1'+\beta_2')v^2_{0,\alpha} + \mathrm{Var}\beta_2' v^2_{2,\alpha} + \mathrm{Var}\beta_2' w^2_{2,\alpha} > \mathrm{Var}\beta_1 v^2_{0,\alpha} = \mathrm{Var}\ p_{2,\alpha}(\beta_1,0).$$

An increase in the mean or variance after the introduction of a measure is consistent with gaming but it is also consistent with an allocation of effort to non-gaming activities that are measure specific (multi-tasking). This implies that the evidence of performance outcome increases is not sufficient to conclude that gaming takes place. Changes in correlation, however, can provide evidence of gaming, as we argue next.

The correlation between p_2 and V, before the introduction of measure two is

$$Corr(p_{2,\alpha}(\beta_1,0),V_\alpha(\beta_1,0)) = \frac{Var v^2_{0,\alpha}}{\sqrt{Var v^2_{0,\alpha} Var(v^2_{0,\alpha}+v^2_{1,\alpha})}}$$

The correlation is less than one because performance measure two does not capture the tasks that are specific to measure one. When measure two is introduced in the contract, the agent starts to invest in measure specific effort, and also in measure specific gaming. The correlation changes to

$$Corr(p_{2,\alpha}(\beta_1',\beta_2'),V_\alpha(\beta_1',\beta_2')) = \frac{Var(\beta_1'+\beta_2')v^2_{0,\alpha} + Var\beta_2' v^2_{2,\alpha}}{\sqrt{(Var(\beta_1'+\beta_2')v^2_{0,\alpha} + Var\beta_1' v^2_{1,\alpha} + Var\beta_2' v^2_{2,\alpha})(Var(\beta_1'+\beta_2')v^2_{0,\alpha} + Var\beta_2' v^2_{2,\alpha} + Var\beta_2' w^2_{2,\alpha})}}$$

Proposition 1: A performance measure is gameable if the correlation between the measure and the principal's objective decreases after the introduction of the measure.

Proof: See Appendix.

The correlation between measure two and the true goal can increase or decrease after that measure has been introduced and this will depend on the relative impact of gaming noise and measure specific effort. It is worth considering two benchmark cases. Consider first the case where there are no measure specific actions ($v_{2,\alpha}$=0 for all α) and assume $Varw_{2,\alpha}^2 > Varv_{1,\alpha}^2$. Then, the correlation between measure two and the true goal decreases after that measure has been introduced. The intuition for this finding is that the introduction of the performance measure increases the noisiness of measure two and therefore decreases its predictive power. Consider next the case where there are no gaming margins ($w_{i,\alpha}$=0 for all α). In this case the correlation increases and this is because the agent invests more in measure specific effort which increases the predictive power of the measure (the covariance between V and m_2 increases).

Proposition 1 implies that a sufficient condition to conclude that a measure generates gaming is to test whether the correlation between the measure and the true goal decreases after the measure is introduced. If the correlation stays constant or increases then there could still be gaming but in an amount low relative to the measure specific effort responses. The advantage of this proposition is that it requires from the researcher little information on the incentive contract. In particular, the identification of gaming does not necessitate any knowledge on the incentive weights, which is notoriously difficult to obtain. To conduct that test, one only needs to know when a performance measure is introduced and to observe the true goal and the measure before and after introduction.

Our model formally demonstrates the suspicion that the correlation between a performance measure and value-added is endogenous. Baker (2002) argues that a correlation measure does not tell the incentive designer anything about the gaming strategies available to the agent. Our model not only confirms that point but it also shows that the change in correlation that takes place after the activation of a measure can be used to identify gaming. Note that others have also argued that performance outcomes should change after a measure is activated. For example, Meyer and Gupta (1994) argued that the worthiness of a performance measure degrades after it is activated. The concept of degradation, however, does not suggest clear statistical predictions for how the measure and the true goal should change.

Finally, our model shows that using a correlation measure to identify good performance measures can be misleading. To illustrate this point, assume the principal is considering adding a new performance measure to complement an existing one. Assume there are two candidate measures, measure 2 and 2' that are identical in all respects, but performance measure 2 is more correlated with the principal's objective than measure 2'. Selecting measure 2 on this criterion may be misleading. In fact, it may turn out that the correlation between the true goal and measures 2 drops after its introduction. It is even be possible that measure 2' is more correlated with the true goal, if the correlation is measured after the measure's introduction. Although our model does not provide a method to select performance measures, it suggests that one has to be cautious in using a correlation based selection criterion.

3 Empirical Application to a Government Job Training Program

3.1 Literature Review and Preliminary Evidence

Data on organizations that relate performance outcomes to measures of organizational value are scarce. This kind of data exists for a large federal job training program and for this reason and this program's experience with performance measurement systems it is the focus of our empirical work. Here we describe the organization that we study and the empirical literature examining performance measurement that bears on our study.

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. In the 1970s several influential studies showing the ineffectiveness of this job training prompted Congress to reconsider how job training programs were constituted. Beginning with the Job Training Partnership Act (JTPA) of 1982 and continuing under the legislation that supplanted JTPA, the Workforce Investment Act (WIA) of 1998, the bureaucracy that runs the federal government's most important job training program for the poor has become highly decentralized. Training is conducted by over 600 local job training centers, each enjoying substantial discretion over who they enroll and what types of training they provide their enrollees. By allowing this discretion, Congress hoped that job training administrators would be free to use their expertise in training and their superior knowledge of "conditions on the ground" to provide better training. But in increasing administrators' 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, Congress has sought to provide stronger incentives to promote programmatic objectives by linking financial incentives to measures of program outcomes. Thus, since JTPA's passage, training center budgets have been partly contingent on their performance on explicitly defined measures. Under JTPA—our analysis of the program focuses on the late 1980s—these measures were variants of program participants' employment and wage rates measured at the time the participants "graduated" from their training.

Performance measure validation literature

JTPA's stated goal was to promote increases in the employment and earnings of enrollees (JTPA, Section 106(a)). Numerous studies have attempted to test the ability of such short-term outcome-based measures to capture long-term earnings and employment gains of enrollees. These studies have been conducted using job training data from JTPA, but also from other job training programs that had not been subject to performance-based measurement.

Gay and Borus (1980), Friedlander (1988) and Zornitsky, et al. (1988) conducted their studies of the association of short-run outcomes and long-term employment and earnings gains based on data from job training programs that had no explicit performance measurement backed by financial incentives. Gay and Borus found that the correlation of employment measures and earnings impacts were sometimes negative. In contrast, Friedlander and Zornitsky both report that enrollees who were likely to produce high scores on employment-based performance measures were also likely to generate high earnings and employment impacts. In their studies based on data generated from JTPA, Heckman, Heinrich, and Smith (2002) and Barnow (2000), however, found little evidence that the performance measures and earnings impacts were significantly correlated. Barnow concluded "there is only a weak correspondence between the two measures and that the Department of Labor should avoid making significant rewards or sanctions based on [them]." (Barnow, p. 118) Heckman, Heinrich, and Smith found that the performance measures "are weakly and sometimes perversely, related to long-term impacts." (Heckman et al, p. 778) 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. What is interesting about the above-cited studies for the purpose of our study is that only in programs where performance is uncompensated (Friedlander and Zornitsky et al) have researchers found statistically significant correlation between short-term performance measures and impacts.

Of course this observation is not definitive because we compare studies that are based on different programs and on different methodologies. Some of these studies construct their measures of job training success using data from social experiments, while others construct them by comparing the labor market outcomes of persons who obtained training to outcomes of persons from an artificially constructed control group. An analysis using a consistent methodology and data from a single program subject to exogenous variation in performance measures in an organizational environment that is in other ways unchanging would be more definitive. We describe such an analysis below.

3.2 Test using JTPA Data

In the mid-1980s several years after the program began, the U.S. Department of Labor (DOL) changed the performance measures used to evaluate bureaucratic performance. In the early years of JTPA, performance measures were based on an enrollee's employment status at the date the enrollee officially exited—or graduated—from the program. In the late 1980s and early 1990s, DOL began to de-emphasize measures based on labor market outcomes at the time of graduation in favor of measures based upon outcomes measured 90 days after graduation. By moving to measures that captured labor market outcomes further removed from job training, DOL hoped to encourage training centers to offer more substantive training that would produce longer-lasting impacts on enrollees' skills. DOL required states to implement these new measures, but gave states some leeway in how quickly they were added. Thus, different states made these transitions in different years.[7] Table 1 defines the new performance measures.

We evaluate the relation between the goal of the organization and the new performance measures described in Table 1 before and after their (exogenous[8]) activation.[9] Here, with minor exception, states were taking on additional performance measures during this period; states for the most part had not yet discarded the graduation-based measures.[10] We develop our empirical measures of performance outcomes and programmatic impacts using data from the National JTPA Study (NJS), an experimental study of the effectiveness of JTPA commissioned by DOL and conducted between 1987 and 1989. Sixteen of the organization's roughly 640 job training centers

[7] See Courty and Marschke (2003b) for a description of the performance measures, incentive system, and the reasons for the changes in the performance measures in these years. Courty and Marschke also detail the timing of the performance measure changes by state.

[8] For evidence and an argument that the establishment of performance measures were indeed exogenous to training centers, see Marschke (2003) and Cragg (1997).

[9] Note that we focus in this analysis on the adult side of JTPA, and ignore the smaller youth side.

[10] Fourteen of the sixteen states added one or more of the follow-up measures described in Table 1 during the period we study. Over the same period, two states dropped a cost standard that had rewarded training centers for keeping costs per employment at graduation low. We omit the cost measure from our change analysis because we cannot produce training cost estimates at the enrolee level using our data.

participated in the NJS.[11] The study was conducted using a classical experiment methodology according to which JTPA applicants were randomized into treatment and control groups. The control groups did not receive JTPA training services for at least 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 adults from the set of participants in the NJS. The data contains 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.[12] In addition, the data contains enrolment and graduation 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 at the enrollee level.

We follow the methodology of Heckman, Heinrich, and Smith (2002), who examine the correlation between JTPA's performance measures (the *M* in our model) and the earnings and employment impacts (the *V*) 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. The basic idea of our analysis is that we construct performance outcome estimates and employment and earnings impacts for various subgroups of the sample. We then correlate these subgroup outcomes and impact estimates and examine how the activation changes with the activation of the corresponding measure.

[11] See Doolittle and Traeger (1990) for a description of the implementation of the National JTPA Study, and Bloom et al. (1997) for a detailed description of its results.

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

We first identify for each training centre 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 (see Courty and Marschke, 2003b). We then assign each experimental participant to one of two subsamples based on whether their random assignment date occurred in a program year in which their training agency was evaluated by the performance measure. Without making some strong assumptions, individual-specific earnings and employment impact estimates cannot be constructed from experimental data (see Heckman, 1992, and Heckman, Smith, and Clements, 1997). Instead, following Heckman et al, we construct impact estimates for subgroups based on individual characteristics measured at the point of application. For each subsample, we construct 56 subgroups based on marital status, welfare/AFDC/Food Stamp receipt, race, age, gender, educational attainment, employment status at application, earnings in the year preceding application, and training centre. Thus, if an individual's data are complete he or she appears in our sample 56 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 his/her earnings over the 18 months following their random assignment.[13] 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.[14] 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 2A shows the estimated earnings and employment impacts for many of the subgroups we created. Table 2A shows that the impacts are often small relative to

[13] Following Heckman, Heinrich, and Smith, to limit the influence of outliers, we delete from our sample persons' in the top one percentile of self-reported earnings.

[14] For a comprehensive discussion of the Bloom assumption and of the problem of drop-outs in experimental evaluations more generally, see Heckman, Smith, and Taber (2002).

their standard errors. This is consistent with findings using these data reported elsewhere (Heckman, Heinrich, and Smith).   This exercise produces for each of the three performance measures that we study, earnings and employment impacts for up to 112 subgroups: one set of up to 56 subgroups of enrollees trained in regimes where the performance measure is activated, and another set of up to 56 subgroups of enrollees trained in regimes where the performance measure is not activated.

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 enrolment and graduation dates from agency records. We constructed the enrollee-level follow-up date-based performance outcomes using the enrollee's reported employment hours and wage information from the calendar month containing the graduation date through the calendar month containing the follow-up date (the follow-up date occurs ninety days after the graduation date). In computing the enrollee-level employment rate at follow-up outcome, we considered an enrollee employed at follow-up if he/she showed employment in the third calendar month following graduation. We constructed the average weekly earnings at follow-up outcome by computing the average weekly earnings of all the enrollee's employment spells ongoing in the third month following the graduation month and then summing over all spells. To be consistent with JTPA's definition of the measure, we constructed the earnings outcome only for enrollees who were employed in the third month following graduation. We constructed the average weeks worked outcome by aggregating the number of weeks of employment over the three month follow-up period. Then, for each of the performance measures, we computed the subgroup performance outcomes by averaging the individual performance outcomes within each subgroup, producing up to 56 separate subgroup outcomes for enrollees whose performance on the measures counts toward the training center's award, and another (up to) 56 subgroup outcomes for enrollees whose performance does not count. Table 2B describes the means of the three performance outcomes for selected subgroups in our sample.

3.3 Results

We regress subgroup estimated employment and earnings impacts on their performance outcomes, weighting the regression by the inverse of the Eicker-White standard errors from the impact estimations. In using a regression framework, we are following Heckman, Heinrich, and Smith, but also the performance measure validation literature in accounting (see, e.g., Ittner and Larcker and Banker, Potter, and Srinivasan). Note that this simple regression of V on P yields an estimate of cov(P,V)/var(P). We thus test whether the coefficient on the performance outcome falls with the activation of the corresponding measure. We take a finding that the coefficient falls as evidence that activating a performance measure weakens its association with programmatic impacts and implies gaming. Because we have two impact measures and three outcome measures we have six equations, which we estimate jointly (using a seemingly unrelated regression framework).

Evidence of Gaming

Table 3 shows the results of our estimation. The dependent variables are the estimated subgroup earnings and employment impacts. Each equation contains on the right hand side the subgroup outcome (either the employment rate at follow-up, average weeks worked at follow-up, or average weekly earnings at follow-up) and the outcome interacted with a dummy variable indicating whether the performance measure is activated. (Note that this model deviates from Heckman, Heinrich, and Smith only by the inclusion of the activation dummy variable.) Each regression also contains an intercept and the activation dummy alone, whose coefficient estimates are omitted from the table.

First, note that the coefficient estimates on the performance outcomes are all positive and significant. This suggests that the new performance outcomes and impacts are indeed correlated when the performance outcomes are not awarded. Next, note that the coefficient estimates on the interacted terms are *jointly* significant (the p value of the joint significance test is .0001). This finding alone is consistent with Baker's model of gaming, which implies a change in the correlation between the performance outcome and the goal with the activation of the performance measure. Third, note that in three of the six cases—one case for each of the three performance measures—the coefficient estimate on the interacted term is negative and significant. The drop in the correlation between outcome and impact is consistent with the award triggering gaming activities. The results suggest that each measure is gameable.

Our model predicts that the variance of the measure should rise with activation, both because activation elicits additional efforts and because activation elicits gaming. Table 4 shows the sample variances of the performance outcomes with and without activation of the corresponding performance measures. Our evidence provides some support for the model's prediction. In the case of the average weeks worked at follow-up and average weekly earnings at follow-up measures, the variances increase with activation. In the latter case, the change in variance is statistically significant by conventional significance criteria. The increase in the variance of the average weeks worked at follow-up outcome is marginally significant by conventional significance criteria.

Evidence of Changes in Performance Measure Ranking

Researchers and practitioners have used correlation methods to sort candidate performance measures and to rank them. The gaming model and the evidence of Table 3, however, show that the correlation between outcomes and goals is endogenous. We can also test empirically whether a ranking of performance measures is endogenous. Table 5 shows the explanatory power of each of the three performance measures in impact regressions by whether the measure is activated. The first line of the table shows the slope coefficient estimates and R-squareds of the regressions of earnings impacts on employment rate at follow-up, average weeks worked at follow-up, and average weekly earnings at follow-up outcomes, respectively, in training centre-years where the corresponding performance measure is not activated.[15] The second line shows the R-squared for these regressions using training centre-year data in which the performance measure is activated. The bottom two lines repeat the comparison for the employment impact measure.

Consider the results for the top half of the table which show the effect of activating performance measures on the earnings impact regressions. Note first that the coefficient estimates are all significant and positive when the performance measure is not activated. When they are not activated, a ranking of the performance measures by R-squared places the employment rate at follow-up measure behind both the average weeks worked and average weekly earnings at follow-up measures. When the performance measure is activated, however, while the coefficient estimate for the employment rate at follow-up remains positive and significant, the coefficient estimates for the other measures fall—indeed they become insignificant—along with their R-squareds. Activating the performance measures, therefore, reverses their rankings: after activation, the employment rate at follow-up measure dominates the other two measures.

The bottom half of the table shows the effect of activating performance measures on the employment impact regressions. When the measures are not activated, only in the employment rate at follow-up regression is the slope coefficient estimate significant. When the measures are activated, all coefficient estimates are insignificant, but the R-squared of the employment rate at

[15] The R-squared is a measure of explanatory or predictive power of the performance measure because it shows the fraction of the variation in the impact that is explained by the variation in the performance outcome. Others in the literature have evaluated performance measures by a comparison of R-squareds; see Ittner and Larcker, pp. 14-15. The results shown in Table 5 differ from the earlier results of Table 3 because the earlier results are generated from a single SUR regression. The results of Table 3 are generated from twelve separate regressions: two dependent variables (the two impact measures) crossed with three independent variables (the three performance outcomes) for each of two subsets of the data (persons trained subject to the corresponding performance measure and persons trained absent the corresponding performance measure). We estimated the twelve models separately so as to observe how performance measure activation affected R-squareds.

follow-up regression is higher than in the others. Thus, in the case of employment impacts, activation does not affect the ranking of performance measures.

4 Conclusions

An important lesson from the incentive literature is that explicit incentives may elicit dysfunctional and unintended responses, also known as gaming responses. These responses, however, are typically hidden from the researcher. This paper develops a general approach to identify gaming. We extend Baker 2002's model to show that one can identify gaming by estimating how the correlation between a performance measure and the true goal of the organization changes with the activation of the measure.

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 the 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 conclusive evidence consistent with our hypothesis. We conclude that the new measures were gameable. These findings are corroborated by our previous work that used the specific rules of the performance measurement system to demonstrate gaming; this work showed that training program managers in the JTPA organization strategically time the reporting of their performance outcomes (Courty and Marschke, 2004a).

The paper also contributes to the literature on the implementation of performance measurement (Ittner and Larcker (1998), Banker, Potter, and Srinivasan (2000), Heckman, Heinrich, and Smith (2002)). Our evidence suggests that using a correlation measure to identify good performance measures can be misleading. 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, we show that a ranking of the performance measures according to how correlated they are with the principal's objective can change after the performance measures have been introduced.

Appendix: Proof of Proposition 1

Proposition 1: A performance measure is gameable if the correlation between the measure and the principal's objective decreases after the introduction of the measure.

Proof: The correlation between p_2 and V, before the introduction of the measure can be rewritten as

$$Corr(p_{2,\alpha}(\beta_1,0),V_\alpha((\beta_1,0))) = \frac{1}{\sqrt{1+\frac{Varv_{1,\alpha}^2}{Varv_{0,\alpha}^2}}}$$

Similarly, the correlation after measure 2 is introduced is

$$Corr(p_{2,\alpha}(\beta_1',\beta_2'),V_\alpha(\beta_1',\beta_2')) = \frac{1}{\sqrt{(1+\frac{Var\beta_1'v_{1,\alpha}^2}{Var(\beta_1'+\beta_2')v_{0,\alpha}^2+Var\beta_2'v_{2,\alpha}^2})(1+\frac{Var\beta_2'w_{2,\alpha}^2}{Var(\beta_1'+\beta_2')v_{0,\alpha}^2+Var\beta_2'v_{2,\alpha}^2})}}$$

The correlation decreases after the introduction of measure 2 if

$$Corr(p_{2,\alpha}(\beta_1,0),V_\alpha((\beta_1,0))) > Corr(p_{2,\alpha}(\beta_1',\beta_2'),V_\alpha(\beta_1',\beta_2'))$$

Since $\frac{Varv_{1,\alpha}^2}{Varv_{0,\alpha}^2} > \frac{Var\beta_1'v_{1,\alpha}^2}{Var(\beta_1'+\beta_2')v_{0,\alpha}^2+Var\beta_2'v_{2,\alpha}^2}$, the above inequality implies that

$$\frac{Var\beta_2'w_{2,\alpha}^2}{Var(\beta_1'+\beta_2')v_{0,\alpha}^2+Var\beta_2'v_{2,\alpha}^2} > 0\,,$$

which says that measure 2 is gameable. QED

REFERENCES

Asch, B.J. (1990), ‘Do Incentives Matter? The Case of Navy Recruiters’, *Industrial and Labor Relations Review*, 43, 89S-106S.

Baker, G. P. (1992), ‘Incentive Contracts and Performance Measurement’, *Journal of Political Economy*, 100(3), 598-614.

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

Baker, George, Robert Gibbons and Kevin J. Murphy. (1994) “Subjective Performance Measures in Optimal Incentive Contracts.” *The Quarterly Journal of Economics.* Volume 109, Issue 4, 1125-56.

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

Banker, R., Potter, G., and Srinivasan, D. (2000), “An Empirical Investigation of an Incentive Plan that Includes Nonfinancial Performance Measures,” *The Accounting Review*, 75(1), 65-92

Barnow, B. (2000). “Exploring the Relationship Between Performance Measurement and Program Impact,” *Journal of Policy Analysis and Management*, 19(1), 118-141.

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

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997) “The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study,” *The Journal of Human Resources*, 32(3), 549-576.

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.

Courty, P. and Marschke, G. (2003a). Dynamics of Performance Measurement Systems. *Oxford Review of Economic Policy*. 2003, 19 (2), 268-84.

Courty, P. and Marschke, G. (2003b). Performance Funding in Federal Agencies: A Case Study of a Federal Job Training Program. *Public Budgeting and Finance*. Fall issue (Vol. 23:3). 22-48.

Courty, P., and Marschke, G. (2004). An Empirical Investigation of Gaming Responses to Explicit Performance Incentives, *Journal of Labor Economics*, 22(1), 23-56.

Cragg, M. (1997). Performance incentives in the public sector: Evidence from the Job Training Partnership Act. *Journal of Law, Economics and Organization* 13 (April), 147–68.

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

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

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

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

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

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

Gibbs, Michael, Kenneth Merchant, Wim Van der Stede, and Mark Vargus. (2004) 'Performance Measure Properties and Incentives.' Chicago GSB Mimeo.

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

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.

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

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

Heckman, J. J., Smith, J., and Taber, C. (2002), `Accounting for Dropouts.’ *Review of Economics and Statistics*.

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

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

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

Marschke, G. (2003) "Performance Incentives and Organizational Behavior: Evidence from a Federal Bureaucracy," Manuscript, University at Albany, State University of New York.

Meyer, M. W. and Gupta, V. (1994). The Performance Paradox. *Research in Organizational Behavior*, 16, 309-369

Oettinger, G. (2002), “The Effect of Nonlinear Incentives on Performance: Evidence from ‘Econ 101’”. *The Review of Economics and Statistics*, 84(3), 509-17.

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

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

van Praag, M. and Cools, K. (2001) “Performance Measure Selection: Noise Reduction and Goal Alignment.” Manuscript, University of Amsterdam.

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

Revised JTPA Performance Measures for JTPA's Adult Program

Performance Measure	*Description*
Employment Rate at Follow-up	Fraction of graduates who were employed at 13 weeks after graduation
Average Weekly Earnings at Follow-up	Average weekly wage of graduates who were employed 13 weeks after graduation
Average Weeks Worked by Follow-up	Average number of weeks worked by graduates in 13 weeks following graduation

Notes:

1. The date of graduation is the date the enrollee officially exits training. A graduate is an enrollee after he/she has officially exited training.
2. All measures are calculated over the year's *graduate* population. Therefore, the average follow-up weekly earnings for 1987 was calculated using earnings at follow-up for the graduates who graduated in 1987, even if their follow-up period extended into 1988. Likewise, persons who graduated in 1986 were not included in the 1987 measure, even if their follow-up period extended into 1987.

Table 2A
Experimental Impacts By Subgroup

Subgroup	18 Month Earnings Impacts ($)	18 Month Employment Impacts (months)
	Receiving Food Stamps	
No	490.823	0.117
	(273.314)	(0.191)
Yes	269.646	0.298
	(281.878	(0.235)
	Gender	
Male	46.236	0.099
	(335.312)	(0.219)
Female	574.451	0.240
	(225.046)	(0.200)
	Highest grade completed	
< 10 yrs	508.109	0.458
	(466.333)	(0.371)
10-11 yrs	455.366	0.314
	(406.520)	(0.323)
12 yrs	528.377	0.104
	(314.432)	(0.237)
13-15 yrs	115.078	-0.312
	(597.620)	(0.399)
> 15 yrs	-394.491	0.735
	(1210.037)	(0.718)
	Race	
White	513.505	0.070
	(261.964)	(0.192)
Black	373.860	0.362
	(358.973)	(0.282)
Hispanic	-330.982	0.212
	(629.107)	(0.458)
Other	85.816	1.008
	(1156.137)	(0.846)
	Age	
22-29 yrs	582.795	0.436
	(303.606)	(0.220)
30-39 yrs	337.990	0.208
	(340.071)	(0.253)
40-49 yrs	167.248	-0.220
	(528.506)	(0.414)
50-54 yrs	-829.954	-1.722
	(1043.039)	(0.881)
> 54 yrs	588.204	0.197
	(765.385)	(0.758)
	Employment status at time of application	
Currently employed	1037.772	0.191
	(483.793)	(0.324)
Last employed 0-2 months ago	196.212	0.009
	(470.613)	(0.318)
Last employed 3-5 months ago	-413.393	-0.086
	(562.806)	(0.382)
Last employed 6-8 months ago	607.643	0.440
	(746.237)	(0.544)
Last employed 9-11 months ago	1952.2044	0.805
	(949.928)	(0.695)
Last employed > 11 months ago	463.599	0.785
	(459.239)	(0.366)
Never employed	291.431	0.327
	(462.171)	(0.426)

Table 2A (continued)
Experimental Impacts By Subgroup

Subgroup	18 Months Earnings Impacts ($)	18 Months Earnings Impacts (months)
	Training Center	
Corpus Christi, TX	-847.637	-0.295
	(731.918)	(0.531)
Cedar Rapids, IA	1057.610	-0.282
	(1295.772)	(0.997)
Coosa Valley, GA	1271.657	0.666
	(633.922)	(0.471)
Heartland, FL	1005.343	1.363
	(1270.066)	(0.942)
Fort Wayne, IN	-417.937	-0.644
	(505.544)	(0.369)
Jersey City, NJ	191.213	-0.277
	(985.257)	(0.729)
Jackson, MS	1974.543	1.633
	(747.497)	(0.557)
Larimer, CO	-12.400	0.490
	(874.319)	(0.628)
Decatur, IL	14.810	0.171
	(1171.069)	(0.783)
Northwest, MN	-2272.185	-2.287
	(1170.215)	(0.949)
Butte, MT	-913.574	-0.690
	(1098.605)	(0.801)
Omaha, NE	1185.606	1.335
	(662.263)	(0.576)
Marion County, OH	968.491	0.151
	(667.529)	(0.549)
Oakland, CA	-1034.381	-0.020
	(850.677)	(0.593)
Providence, RI	966.893	0.830
	(844.981)	(0.618)
Springfield, MO	378.675	-0.305
	(710.688)	(0.481)
	Earnings at time of application	
$0-$3,000	497.049	0.378
	(245.126)	(0.204)
$3,000-$6,000	487.801	0.075
	(521.988)	(0.347)
$6,000-$9,000	-439.713	-0.295
	(704.655)	(0.440)
$9,000-$12,000	842.508	-0.358
	(1091.196)	(0.635)
$12,000-$15,000	1569.974	0.636
	(1675.107)	(0.907)
>$15,000	-613.327	-0.211
	(2086.843)	(1.070)

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. Subgroups created using AFDC receipt, marital status, and family size excluded for space considerations.

Table 2B
Mean Performance Outcomes By Subgroup

Subgroup	Employment Rate at Follow-up	Average Weeks Worked at Follow-up (weeks)	Average Weekly Earnings at Follow-up ($)
		Receiving Food Stamps	
No	0.574	8.868	230.247
	(0.495)	(5.657)	(119.805)
Yes	0.496	7.113	207.264
	(0.500)	(6.079)	(118.138)
		Gender	
Male	0.531	8.393	8.393
	(0.499)	(5.820)	(5.820)
Female	0.520	7.918	7.918
	(0.500)	(5.960)	(5.960)
		Highest grade completed	
< 10 yrs	0.469	7.388	207.738
	(0.499)	(6.024)	(114.193)
10-11 yrs	0.484	7.419	214.441
	(0.500)	(6.038)	(104.217)
12 yrs	0.545	8.414	219.978
	(0.498)	(5.822)	(121.682)
13-15 yrs	0.572	8.883	242.324
	(0.495)	(5.659)	(135.507)
> 15 yrs	0.692	9.764	250.349
	(0.463)	(5.239)	(127.915)
		Race	
White	0.566	8.601	221.835
	(0.499)	(5.788)	(122.813)
Black	0.467	7.267	227.086
	(0.499)	(6.031)	(123.865)
Hispanic	0.468	7.801	207.418
	(0.499)	(5.966)	(100.502)
Other	0.525	8.186	237.655
	(0.500)	(5.784)	(110.805)
		Age	
22-29 yrs	0.525	8.194	220.764
	(0.499)	(5.898)	(105.620)
30-39 yrs	0.525	8.165	228.466
	(0.499)	(5.858)	(124.000)
40-49 yrs	0.522	7.939	223.282
	(0.500)	(5.969)	(144.536)
50-54 yrs	0.500	7..558	219.755
	(0.501)	(6.053)	(172.393)
> 54 yrs	0.553	8.068	161.767
	(0.498)	(6.028)	(102.369)
		Employment status at time of application	
Currently employed	0.666	10.120	221.082
	(0.472)	(5.054)	(111.279)
Last employed 0-2 months ago	0.594	8.949	234.982
	(0.491)	(5.587)	(143.728)
Last employed 3-5 months ago	0.552	8.658	231.269
	(0.498)	(5.660)	(113.235)
Last employed 6-8 months ago	0.522	8.168	229.595
	(0.500)	(5.856)	(108.832)
Last employed 9-11 months ago	0.525	8.098	217.886
	(0.500)	(5.922)	(101.381)
Last employed > 11 months ago	0.435	6.721	209.313
	(0.496)	(6.108)	(125.440)
Never employed	0.392	6.302	190.362
	(0.489)	(6.189)	(102.424)

Table 2B
Mean Performance Outcomes By Subgroup

Subgroup	Employment Rate at Follow-up	Average Weeks Worked at Follow-up (weeks)	Average Weekly Earnings at Follow-up ($)
		Training Center	
Corpus Christi, TX	0.499	8.500	193.378
	(0.501)	(5.780)	(111.705)
Cedar Rapids, IA	0.559	8.479	235.631
	(0.498)	(5.911)	(179.038)
Coosa Valley, GA	0.581	8.168	232.552
	(0.494)	(5.847)	(134.171)
Heartland, FL	0.504	7.779	207.493
	(0.502)	(5.502)	(82.616)
Fort Wayne, IN	0.658	9.873	216.703
	(0.475)	(5.213)	(99.946)
Jersey City, NJ	0.411	6.244	268.118
	(0.493)	(6.062)	(112.040)
Jackson, MS	0.578	8.291	210.692
	(0.494)	(5.737)	(134.768)
Larimer, CO	0.509	8.553	218.458
	(0.500)	(5.855)	(122.377)
Decatur, IL	0.643	9.562	247.448
	(0.480)	(5.237)	(149.839)
Northwest, MN	0.571	8.736	222.843
	(0.497)	(5.938)	(90.900)
Butte, MT	0.509	7.923	234.589
	(0.501)	(5.969)	(178.290)
Omaha, NE	0.508	7.279	197.955
	(0.500)	(6.025)	(91.602)
Marion County, OH	0.411	6.451	197.563
	(0.492)	(6.190)	(107.089)
Oakland, CA	0.393	7.120	271.745
	(0.489)	(6.121)	(125.767)
Providence, RI	0.369	5.977	249.088
	(0.483)	(6.137)	(96.094)
Springfield, MO	0.702	10.250	209.348
	(0.458)	(4.871)	(98.995)
		Earnings at time of application	
$0-$3,000	0.467	7.286	210.294
	(0.499)	(6.042)	(123.507)
$3,000-$6,000	0.601	9.163	223.564
	(0.490)	(5.503)	(114.387)
$6,000-$9,000	0.665	9.966	240.109
	(0.473)	(5.108)	(112.070)
$9,000-$12,000	0.642	10.130	265.643
	(0.480)	(4.968)	(113.000)
$12,000-$15,000	0.609	9.711	293.612
	(0.490)	(5.497)	(98.025)
>$15,000	0.694	10.078	345.931
	(0.463)	(5.233)	(192.342)

Notes: Standard deviations are reported in parentheses. The performance outcome means are reported from 10746 adult experimental participants who report a valid earnings figure in each of the 18 months after random assignment. Subgroups created using AFDC receipt, marital status, and family size excluded for space considerations.

Table 3

Outcome-Impact (SUR) Regressions

Coefficient	Employment Rate at Follow-up	Average Weeks Worked at Follow-up	Average Weekly Earnings at Follow-up
Dependent Variable = 18 Month Earnings Impact			
Performance outcome	1478.014	67.570	2.548
	(6.16)	(6.44)	(6.38)
Performance outcome X Activation Dummy	924.906	-80.582	-3.708
	(0.99)	(-2.22)	(-2.05)
Dependent Variable = 18 Month Employment Impact			
Performance Outcome	1.009	0.031	0.001
	(5.71)	(4.16)	(4.11)
Performance Outcome X Activation Dummy	-1.616	-0.040	-0.002
	(-2.31)	(-1.72)	(-1.51)
R^2		0.3947	

Notes: T statistics in parentheses. 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. Regressions are weighted by the inverse of the Eicker-White standard errors from the impact estimations. Earnings are trimmed in construction of impact estimates.

Table 4
Test of Performance Measure Activation on Variance of Performance Outcome

Performance Measure Activated	Employment Rate at Follow-Up	Average Weeks Worked at Follow-up	Average Weekly Earnings at Follow-up
		Sample Variance	
No	0.250	34.287	13958.75
Yes	0.249	36.233	16318.43
F Test of equal variances (p value)*	0.505	0.105	0.003

*F test: $H_o : \sigma^2_{Yes} = \sigma^2_{No},\ H_a : \sigma^2_{Yes} > \sigma^2_{No}$

Table 5
R^2's and P Values and from Least Square Regressions of Impacts on Outcomes
By Whether Performance Measure Activated*

		Employment Rate at Follow-Up			Average Weeks Worked at Follow-up			Average Weekly Earnings at Follow-up		
Dependent Variable	Performance Measure Activated	Coef. Est.	P Value	R^2	Coef. Est.	P Value	R^2	Coef. Est.	P Value	R^2
Earnings Impact	No	1167.20	0.0008	0.215	72.40	<0.0001	0.370	2.84	<0.0001	0.383
	Yes	1039.18	0.0003	0.270	-16.51	0.6756	0.004	-0.47	0.7336	0.003
Employment Impact	No	0.56	0.0204	0.109	0.01	0.3632	0.018	0.00	0.1761	0.039
	Yes	0.11	0.6244	0.006	0.00	0.9613	0.000	0.00	0.8498	0.001

* This table describes the slope coefficient estimates and R^2's of 12 regressions of impacts on performance outcomes. For each of the six outcome-impact combinations, we perform two regressions: one for individuals trained in regimes with the corresponding performance measure activated and one for individuals without the performance measure activated. Regressions are weighted by the inverse of the Eicker-White standard errors from the impact estimations. Earnings are trimmed in construction of impact estimates.