

---

# Benchmarking Performance

Pascal Courty  
Department of Economics  
European University Institute

Gerald Marschke  
Department of Public Administration & Policy  
and Department of Economics  
University at Albany, SUNY

## Abstract

*An important challenge in the design of performance measurement, accountability, and incentive systems is the establishment of relevant benchmark levels of performance, also known as performance standards. We review information economics literature and draw simple lessons for the construction of performance standards. We demonstrate the relevance of these lessons in the context of a job training government organization.*

## Introduction

The construction of performance standards is central to the legitimacy and success of performance measurement systems. Performance standards establish benchmarks to guide the evaluation of actual performance and to construct measures of value-added. Performing above or below the standard can have important short-term consequences (rewards, budget revision) as well as long term ones (promotion, structural reorganization).

In this paper, we review the information economics literature (Prendergast 1999, Dixit 2002) and draw simple lessons for the construction of performance standards. We compare in light of these lessons different methods that are commonly used to construct performance standards. In the second part of the paper, we investigate the relevance of these lessons in the context of the performance measurement system of the large government job training organizations created under the Job Training Partnership

Act (JTPA) and the Workforce Investment Act (WIA). We review the large literature on the design and impact of performance incentives in JTPA and WIA (Barnow, 2000, Courty and Marschke, 2004, Heckman, Heinrich, and Smith 2002, and Heinrich, this issue) focusing on the issue of performance benchmarking. We ask whether the design of performance standards in these two organizations is consistent with basic principles from information economics.

There is large literature on performance management in public organizations. See for example Bouckaert (1993), Wholey and Hatry (1992), Kravchuk and Schack (1996), and Wholey (1999). This literature covers many issues related to performance management, including the administration of performance measurement, the selection of performance measures, timing of performance measurement, and also performance benchmarking. (See especially Hatry, 1999, and the papers cited therein; Stiefel, Rubenstein, and Schwartz, 1999; Brooks, 2000; Rubenstein, Schwartz, and Stiefel, 2003.) Our contribution to this literature is to focus exclusively on the issue of performance benchmarking, and to offer a rigorous economic framework, based on information economics and contract theory. We follow throughout the paper an economic approach. We focus on the role that performance standards play as part of mechanism to provide performance incentives. Our premise will be that the method used to construct performance standards can influence the internal efficiency of organizations by changing the way agents behave.

Obviously, the exercise of assessing performance plays other roles in organizations. For example performance standards are used to learn about best practices and to communicate to outside constituencies legitimate information about organizational value-added. Similarly, we recognize that in reality political concerns as well as ethical values influence the construction of performance standards. We discuss these issues only to the extent that they matter in our case study.

### **Some Economics of Performance Standards**

Some notation will clarify the discussion. Following the economic literature, we call the party who designs the measurement system the principal and the party whose performance is measured the agent. We call the measured outcome the performance outcome,

which we denote  $P$ , the benchmark level of performance the performance standard, which we denote  $P_0$ , and the difference between the performance outcome and the performance standard, excess performance or value added, which we denote  $\Delta P$ . Thus,

$$\Delta P = P - P_0.$$

We assume that the agent has some control over the performance outcome, and following the economics literature, effort constitutes the agent's choices and exertions that influence the performance outcome. We denote the effort choice  $e$  and assume that higher effort levels increase the performance outcome, that is,  $P(e)$  increases with  $e$ . In the simplest formulation, performance is equal to effort  $P=e$ , and value added is equal to  $\Delta P=e-P_0$ .

To set ideas, we give an economic rationale for how to establish the absolute level of the performance standard. Assume the performance standard determines the level of acceptable performance below which sanctions are imposed and above which rewards are given. By increasing the standard, the principal increases performance incentives because the agent has to supply more effort to meet the standard and avoid sanctions. There is a limit, however, to how much the principal can increase the standard because the agent may give up and quit if the standard is too high. To capture this idea we assume that the agent receives a level of compensation that is independent of performance and we define the level of effort that one would expect for that base compensation as  $e_0$ . Prevailing competitive forces determine this level of effort. In other words, it is the amount of effort a representative agent would expect to exert for the base level of compensation. Then, the performance standard is set at the level of performance that occurs when the agent provides the competitive level of effort,

$$P_0 = P(e_0).$$

The rationale behind setting the performance standard at this level is that if the performance standard were set above  $P(e_0)$  then the principal would be unable to attract and retain the agent. On the other hand, if the performance standard were set below  $P(e_0)$  then the principal would be over compensating the agent.

---

To illustrate this definition of the performance standard, consider a simple manufacturing production example. We use this example because manual work constitutes the occupational class where performance benchmarking has been first used in a systematic way. (Taylor, 1911, was perhaps its most famous early proponent. His ideas arose from his experience as a machinist in a steel plant). Performance benchmarking is still common in production manufacturing. This choice of example is without loss of generality, as we will argue that the problems that arise with the construction of performance standards in public organizations do not fundamentally differ from the ones that arise in production manufacturing.

Suppose a manual worker in a factory who is paid  $w$  per hour. The wage is paid independently of the level of worker performance. In addition to the fixed wage, the principal may wish to reward the worker for superior performance and impose sanctions for inferior performance. The number of pieces the worker produces per hour is by itself insufficient to assess whether to reward or sanction the agent. One way to address this question is to conduct time-and-motion studies to establish a benchmark level of performance, or hourly rate,

$$P_0 = e_0(w)$$

that a representative worker earning  $w$  would achieve and use this benchmark to evaluate actual performance. The principal actually assesses the level of performance that occurs under competitive effort and uses this information to set the performance standard. The difference between the worker's performance and the performance standard is used as a measure of value-added. Under that interpretation, value-added corresponds to what the agent adds, because of superior effort, to what we would expect to prevail in the market.

This method of establishing the performance standard requires estimating the production technology available to the agent, that is, the relation between effort and outcomes. Once this relationship is understood, it is possible to infer the agent's excess effort relative to the competitive level of effort. Counter-factual experiments such as time-and-motion studies, however, are practical only in a few occupations that typically involve manual work. Most work situations, however, involve non-manual work, complex group

interactions, and non-standardized outputs, making experimental studies to construct counter-factual performance benchmarks very costly.

These complications mean that the methods typically used to construct performance standards are imperfect. Real world methods balance the cost of establishing accurate standards and the expected return for the organization from assessing value-added more precisely. We will review these methods at the end of this section and consider their pros and cons. First, however, we review a set of generic problems that any construction method must address.

### **Leveling the Playing Field**

So far, we have considered the case of a principal who manages a single agent working in a single environment. It often happens, however, that the agent works in multiple environments or that the principal manages multiple agents who face different work conditions. To illustrate, we return to the time-and-motion study example presented above and we assume that there are multiple workers who are assigned to different machines. Assume the machines vary in their productivity and that each machine's productivity is known to both the principal and the agent. Assume that performance outcome with machine  $k$  when effort is  $e$  is

$$P(e)=ke,$$

where  $k>0$ . Assume also that the average machine exhibits unit productivity ( $E_k=1$ ) and that the agent is allowed to refuse to work on a machine. If this is the case the principal must factor the difference in marginal products of effort across agents into the determination of the performance standard. In fact, if the principal sets the same performance standard for all machines, say  $P_0=e_0$ , then the agent will only agree to work with machines with  $k>1$  and will refuse to work with machines with  $k<1$ . The point is that time-and-motion studies have to be conducted in each work environment to control for the special circumstances of the environment that are commonly observed by the principal and the agent.

### **Insurance and Uncontrollable Risk**

---

The time-and-motion study example presented above assumes that there is no uncertainty about the worker's performance outcome. Consider a more realistic example where the worker produces a number of pieces that depends on his or her effort and also on some outside shock, for example, a power outage that slows production. The performance outcome is now equal to  $P+\varepsilon$  and the worker's excess performance is

$$\Delta P=(P+\varepsilon)-P_0,$$

where  $\varepsilon$  is a mean zero random variable. Setting the standard at a level that does not take into account individual circumstances, implies that a worker who supplies the effort level that is required to achieve the performance standard will sometime over-perform and other times under-perform the performance standard since value added is equal to

$$\Delta P=\varepsilon$$

when  $e=e_0$ . Outside shocks do not influence the worker's choice of effort because additional effort still increases expected excess performance. Outside shocks, however, do change realized excess performance and therefore the workers' compensation. The worker will achieve the same level of excess performance on average but there will be variations around that level. Although a risk neutral worker will not suffer any disutility from this variation in compensation, a risk-averse worker will and this establishes a first economic rationale to construct as accurate a performance standard as possible. The logic is that more accurate standards reduce compensation risk thereby increasing the agents' welfare. Lowering the risk faced by the agent is good for the principal as it means that he does not have to offer the agent a higher wage to compensate for risk-bearing. This suggests that the worthiness of performance standards will depend on their ability to control for outside risk.

The key issue is to identify the source of controllable versus uncontrollable factors. Carefully assessing the factors that influence the performance outcome and that are not within the agent's control is essential. It is important to hold the worker responsible for effort,  $e$ , but not for outside shocks,  $\varepsilon$ . Yet the principal does not want to discount factors that are within the agent's control. For example, the principal does not want to lower the performance standard in the

event the workers' machine breaks down if the agent might have anticipated and prevented the breakdown.

Risk aversion is not the only reason why the agent may experience disutility from performance standards that do not control for outside risks. Another reason, which is closely related to risk aversion although different, is a concern for fairness. Under risk aversion, the main issue is income variability while under fairness other considerations, such as inter-personal comparisons, may also matter. For example, the agent may experience more disutility from an idiosyncratic shock that lowers only her performance and not the performance of her co-workers than from a group shock that lowers all workers' performances because only the former generates different treatments amongst individuals who have essentially behaved identically. Similarly as under risk aversion, the prediction under the fairness concern is that the principal benefits from discounting factors that are outside the worker's control.

### **Hidden Information, Adverse Selection and Distortions**

We have shown in the previous section that in constructing performance standards, the principal may want to take into account information about shocks that influence the performance outcome and that are outside the agent's control. These shocks are observed only after the agent has chosen effort implying that they do not influence this choice. We now consider a different kind of information that plays an important role in the construction of performance standards. This information is observed only by the agent, and not by the principal, and it is observed before the agent chooses his/her level of effort. For example, assume that when the agent is assigned to a new machine, s/he alone knows the productivity of that machine and can use this information to make his/her effort choice. We say that the information is privately known by the agent and we characterize such situations as hidden information.

Hidden information further complicates the problem of setting standards. To illustrate, consider our job training program application. As explained below, the training efforts of caseworkers in this program are evaluated according to, for example, the rate at which their trainees secure jobs. Caseworkers observe relevant information about applicants on the likely success of training investments (e.g. personal motivation and financial situation). Based

on this information, the caseworkers can predict how likely the applicant is to obtain employment by the end of training. Assume furthermore that those applicants who are more likely to perform well on the measure are not necessarily those who benefit most from training. Indeed, some applicants to the job training program may be highly likely to obtain employment on their own. As a result the caseworkers may over-invest in easy-to-serve applicants and under-invest in hard-to-serve ones and it may be impossible for the principal to correct these investment distortions. For example, a given effort level  $e$  could produce performance outcome

$$P=e+h$$

where  $h<0$  if a hard-to-serve participant is enrolled and  $h>0$  if an easy-to-serve participant is enrolled. We denote hidden information by  $h$  to distinguish this kind of information from the publicly known information  $k$ . If the caseworker observes the applicants' type,  $s/he$  has an incentive to enroll only easy-to-serve applicants because they produce higher outcomes.

In fact, the only way the principal could try to correct these distortions would be by controlling for the type of applicants who have been served, rescaling upward the performance of those agents who have enrolled a larger fraction of hard-to-serve applicants. But by assumption only the agent knows this information. If the principal were to ask the agent what type of participants  $s/he$  has enrolled, the agent would report enrolling only hard-to-serve enrollees and the principal would have no way to verify that the agent is telling the truth.

One may argue that the principal could correct these distortions by developing a specific measure to target the hard-to-serve population, using a variable that can be measured, like welfare reciprocity, as a proxy for hard-to-serve. The assumption would be that welfare recipients are harder to serve, because they require higher investments for equal outcomes. The principal may want to set a lower performance standard for welfare participants, for example,

$$P_0'=e_0-h$$

where  $P_0'<P_0$  by construction. (1) But all welfare recipients are not identical; some are easier-to-serve than others and the agent

observes this information. Again, the agent will select a non-representative sample of welfare recipients. This implies that the principal has corrected some distortions because the agent's attention is now focused on a more desirable target population but the agent will still select those applicants who are the easiest to serve within the sub-population of welfare recipient applicants.

Note that a slightly different problem from hidden information, known as adverse selection in the economic literature, occurs when there are multiple agents who are privately informed. In our example, it could be the case that different agents face different costs, which they privately observe, of meeting the standard. In the job training case, this happens when there are multiple caseworkers who face different eligible populations and when the caseworkers privately observe this information. The distinction between adverse selection and hidden information has to do with the point in time when the agent becomes privately informed. Under hidden information, the agent becomes privately informed after agreeing to the contract while under adverse selection the agent is informed before agreeing to the contract. As a consequence, adverse selection introduces the possibility that the agent's private information will influence the agent's decision to accept the contract or not.

To illustrate, assume that the principal offers all agents the option to run special programs that are only for hard to serve populations. The principal lowers the standards for these special programs and using our terminology this would constitute a new optional contract. The agent agrees or declines to participate. The agents will choose to run such programs on the basis of their private information about the population they face. Presumably, the agents who face the best chances to meet the lower performance standards will decide to run such programs but these agents may not be those who generate the highest return from the principal's perspective. The agent's selection rule is adverse to the principal when it does not correspond to the rule that the principal would use had the principal had the same information as the agent.

### **Dynamic Issues**

Measurement systems are often changed from time to time. There are many reasons why the principal may update performance standards. First, the principal may want to set low standards when a

new performance measure is introduced to give the agent time to adjust to the change. Second, the principal may correct performance measurement systems as s/he acquires new information about the effectiveness of different measurement schemes. Third, the principal may revise the standard to account for changes in the environment or in the production technology.

The agent will take into account the possibility of future changes and most importantly the fact that current performance outcomes may be used in setting future standards. Therefore, the dynamics of performance standard changes, and the policies or absence of policies that guide these changes, may trigger different behavioral responses.

Consider an incentive scheme that is revised over time following a rule that increases the performance standards when excess performance is positive. Such a rule implicitly exists in organizations that use past agent performance to estimate the production function and set standards for the present. Assume the agent systematically outperforms the standard and the principal consequently decides to increase it. It could be the case that the agent was outperforming the standard because the agent was exerting exceptional effort. The agent will then anticipate that current performance influences future standards. The natural response to such a rule is to stop supplying high effort because it increases the standard in the future. Thus a simple static view of incentive systems may fail to capture some behavioral responses that arise only when one considers the dynamic nature of performance measurement.

This phenomenon is known as the ratchet effect in the economic literature and it has two tenets. First, the principal increases the standard for those agents who are superior performers. Second, agents anticipate this possibility and may be reluctant to supply superior effort because they are afraid that this could trigger increases in their standards. Fear of the ratchet effect eliminates the incentives that a measurement system is supposed to introduce. The agent's belief about the principal's policies regarding future standards will largely determine his/her behavior and the success of the incentive system.

To eliminate the ratchet effect, the agent must trust that the principal will not change the standard. Trust is more likely to develop under repeated interactions when the principal can create a reputation for not renegeing on the contract. Another way the principal can eliminate the ratchet effect is by committing to never change a standard or more realistically by committing to very strict rules for changing the standard. Such commitment will eliminate the fear of the ratchet effect and reinforce incentives for effort.

### **Overview of Construction Approaches**

We review different performance standard construction methods and discuss the environments where these methods are likely to work well.

#### **Solution 1: Estimation of Production Function**

A performance standard can be based on an estimation of the production function. As we mentioned above, it is sometimes possible to establish a standard through experimentation (e.g. time-and-motion studies for manual workers). Alternatively, the production function can be assessed through statistical methods. Such an approach is valid only for production processes that are stable over time and across environments.

#### **Solution 2: Past Performance**

This solution is really an application of the previous method where past performance outcomes are used to construct estimates of the production function. There are two problems with this method. The main problem is that this method introduces a ratchet effect because higher performance outcomes increase future standards. Another problem is that this method will not work well in non-stationary environments where the production technology is subject to transient shocks.

#### **Solution 3: Relative Performance Evaluation (RPE)**

RPE is possible when the principal manages multiple agents. RPE can take many forms. In one form, the principal ranks the agent's performance as in a tournament. Alternatively, the principal compares the agent's performance to the average performance among all agents who perform the same work. RPE works well for insurance purposes because it controls for shocks that are common to all agents. In this way, the model provides a rationale for benchmarking by

---

comparing performance across similar workers/agencies, as called for by some public administration scholars—see Hatry, 1999. One potentially important problem with RPE, however, is that it may exacerbate competition and may also result in negative investments (e.g., sabotage, monitoring others).

#### **Solution 4: Negotiation of Standard**

Under this method, the principal and the agent agree on the performance standard. This approach may be the only solution in the absence of objective information on the production function and when RPE is inappropriate, say, because it discourages cooperation among agents. The problem with this method is that it may be hard for the principal to commit not to increase the standard in the event of excess performance. As a result, it will work well only in environments where the agent trusts the principal. Another problem with this method is that the resulting performance standard may be more a function of the relative bargaining ability of the two parties as opposed to principles of setting effective performance standards.

#### **Application to a federal job training program**

In the U.S. government's primary job training program for the economically disadvantaged, a performance measurement system has existed for over two decades. During this time, individual providers of government training have been evaluated by their performance relative to specific, numerical standards. Congress has legislated important changes to how these numerical standards have been formulated. This section interprets the standards in this federal job training program using the framework set forth above.

Federal involvement in job training for the economically disadvantaged began in the Kennedy administration and has since been modified by a series of congressional acts, the most recent passed in 1998. In 1982, the Job Training Partnership Act transformed the federal agency that administered the program in two important ways. First, the program under JTPA became highly decentralized: more than 620 semi-autonomous sub-state training centers administered the program with significant discretion over who to admit and how to conduct the training. Second, and most important for this study, in an attempt to improve job training outcomes, the federal government began to evaluate training centers against numerical standards of performance. In addition, as an incentive, training centers that performed well relative to these

standards received modest budgetary increases. (For detailed descriptions of JTPA and its performance measurement system see, for example, Johnston, 1987, Courty and Marschke, 2003, and Dickinson et al., 1988.)

Under JTPA, Congress, the U.S. Department of Labor (DOL), and state authorities shared in designing and implementing the program's incentive policies. However different their motivations, in this paper, these entities collectively constitute the *principal* in the JTPA organizational structure. The act delegated to the training centers the business of enrolling, training, and finding employment for the program's clients. For the purpose of discussion, these semi-autonomous training agencies individually constitute the *agents*. (2)

The act directed training centers to provide services to help the economically disadvantaged to develop skills that would enable them to obtain employment and increase their earnings, to reduce welfare dependency among the eligible population, and to "contribute to occupational development, upward mobility, development of new careers, and overcoming sex-stereotyping in occupations traditional for the other sex." While the act enumerated a number of goals, it only directed DOL to establish an incentive-backed performance measurement scheme to encourage the first goal listed above. That is, the act required DOL to formulate performance measures to evaluate each training center's success at maximizing its return on human capital investment (JTPA, section 106(a)).

Table 1 shows the performance measures JTPA training centers faced between 1984 and the program's cessation in 2000. For example, an important measure early in the program was the *employment rate at termination*, computed as a fraction of enrollees who were employed on the date they officially completed--or were terminated from--the program. Additional measures were based on enrollees' wage rates at training end. Note that most measures capture enrollees' labor market status shortly after they receive training.

### **Determining the Base Level of Performance**

The first challenge in setting performance standards is to establish the counter-factual level of performance that would occur under a competitive level of effort. Consider for example the

employment rate at termination measure. What employment rate outcome would an agent who supplies a competitive level of effort achieve?

TABLE 1  
National JTPA Performance Measures

Performance Measure	Definition
Employment Rate at Termination	Fraction of trainees employed at termination
Welfare Employment Rate at Termination	Fraction of trainees receiving welfare at date of application who were employed at termination
Average Wage at Termination	Average wage at termination for trainees who were employed at termination
Cost per Employment	Training center's expenditures on adults divided by the number of adults employed at termination
Employment Rate at Follow-up	Fraction of trainees who were employed at 13 weeks after termination
Welfare Employment Rate at Follow-up	Fraction of trainees receiving welfare at date of application who were employed at 13 weeks after termination
Average Weekly Earnings at Follow-up	Average weekly wage of trainees who were employed 13 weeks after termination
Average Weeks Worked by Follow-up	Average number of weeks worked by trainees in 13 weeks following termination
Youth Employment Rate at Termination	Fraction of youth trainees employed at termination
Youth Positive Termination Rate	Fraction of youth trainees who were "positively terminated" (see note 3 below)
Youth Employability Enhancement Rate	Fraction of youth trainees who obtained "employment competencies" (see note 3 below)
Youth Cost per Employment	Training center's year's expenditures on youths divided by the number of youths positively terminated

Notes:

1. The data of termination is the date the enrollee officially exits training. A trainee is an enrollee after he has officially exited training.
2. All measures are calculated over the year's *trainee* population. Therefore, the average follow-up weekly earnings for 1987 were calculated using earnings at follow-up for the trainees who terminated in 1987, even if their follow-up period extended into 1988. Likewise, persons who terminated in 1986 were not included in the 1987 measure, even if their follow-up period extended into 1987.
3. A positive termination was considered either entering un-subsidized employment, attaining youth employment "competencies" (through course-work, training and/or tests in work maturity, basic education, or job-specific skills), entering non-JTPA training, returning to school full-time, or completing a major level of education.



DOL solved that problem by using past performance. In anticipation of JTPA's performance measurement system, DOL collected performance data related to all of JTPA's performance outcomes during the final years of the training program that preceded JTPA. These performance outcomes were used to determine the performance standards in the first year of JTPA. To illustrate, let us ignore for now differences in training environments. We consider a representative training environment. Past performance in that environment gives a distribution of performance. Assume for now that differences in performance outcomes are due only to differences in effort. If DOL believes that only 50 percent of the training centers have supplied at least the competitive level of effort then the performance standard should be set at the 50<sup>th</sup> percentile of the distribution of past performance.

In the JTPA measurement system, the performance standard was set at the outcome produced by the training center at the 25<sup>th</sup> percentile of performance among all training centers nationwide. Thus, DOL evaluated a training center's effort,  $e$ , against an effort level  $e'$ , that corresponded to the effort level of the training center at the 25<sup>th</sup> percentile of system-wide performance. An interpretation of this choice is that 25 percent of the training centers in the previous program were not supplying a competitive level of effort. Using past performance to compute current standards also has dynamic implications that we consider at the end of this section.

### **Leveling the Playing Field**

We have ignored so far differences across training centers. In reality, both the populations from which training centers drew their enrollees and the labor markets for which they trained them varied. For example, some training centers located in relatively depressed labor markets could expect low performance outcomes relative to training centers located in relatively tight labor markets, corresponding to  $k < 1$  in the model. In cases where  $k < 1$  then the performance standard measured at the 25<sup>th</sup> percentile of the distribution of performance was a positively biased estimate of the level of performance that the training center would achieve exerting the competitive level of effort.

DOL recognized this problem and provided states with a method to adjust standards that took into account features of the

training center's environment that may have been correlated with the performance outcome. This method established the 25<sup>th</sup> percentile only as a starting or departure point. For each training center, the method adjusted the departure point by the extent that the training center's characteristics differed from the average training center's characteristics. For example, by taking into account local unemployment rates and other measures of the labor market, the adjustment methodology lowered the employment rate standard for training centers in depressed job markets, compared to training centers in robust ones. A higher unemployment rate indicated a greater difficulty locating jobs for enrollees, thus a decreased marginal product of effort.

Unemployment rates also proxied for the skill levels of the JTPA-eligible persons in the local area served by the training center. Often training centers operating in labor markets with high rates of unemployment also served eligible populations particularly deficient in skills. Thus the DOL adjustment method at least partially adjusted for differences amongst training centers in populations served. A more detailed explanation of the adjustment method is provided in the appendix. In particular, the appendix shows how the DOL method was used to establish the performance standard for the employment rate at termination measure for a representative training center. Table A, for example, reveals that DOL used two variables to level the playing field, i.e., to account for heterogeneity in  $k$  in the model: the local population density and unemployment rate.

### **Risk: Distinguishing Controllable and Uncontrollable Actions**

The adjustment methodology corrected also for the risk generated by the  $\varepsilon$  in the model. While excess performance is still an unbiased estimator of excess effort even in the presence of a random shock (denoted  $\varepsilon$  in the previous section), the presence of this shock introduces noise in the measure of value-added and therefore in the training center's award. Because of risk-averse training staff and planning costs caused by uncertain budgets, it was in the principal's interest in formulating performance standards not only to control for persistent differences across training centers, but also transitory and idiosyncratic shocks ( $\varepsilon$  is composed of these), as well.

After the training center selected and trained an enrollee, her labor market outcome, and therefore her contribution to  $P$ , was at least partly influenced by the availability of jobs in the local labor market. The year-to-year variations in job availability cannot be anticipated by training centers. In addition to capturing permanent aspects of a local labor market, measures in the adjustment model like the year's local unemployment rate, captured transitory shocks (such as factory closings). By adjusting a training center's standards for the local unemployment rate for the year in which its performance was measured, the adjustment method reduced the variance in unpredictable changes in the environment. This constitutes an important advantage of the DOL adjustment model and the decision to abandon this model in WIA has generated concerns (Heinrich, this issue).

### **Hidden Information: Cream-Skimming and Quick-Fixes**

In addition to being noisy, JTPA's performance standards were vulnerable to manipulation by the agent. One way to increase  $P-P_0$  and thus the award was to increase effort. Another way to increase  $P-P_0$  that required no additional effort, however, was to select among the eligible applicants only the high-h types. That is, training centers might enroll persons who would produce high employment rates and earnings, even in the absence of training. This behavior has been called cream-skimming (see, e.g. Anderson, Burkhauser, and Raymond, 1993; Cragg, 1997; Heckman and Smith, 2003; and Heckman, Smith, and Taber, 1996).

To prevent cream-skimming, DOL adjusted standards for the effects of the characteristics of enrollees on  $P$ . As the appendix to the paper illustrates, the adjustment method compensated training centers for enrolling persons such as the handicapped who tended to lower post-training employment rates and earnings outcomes. Training centers that enrolled lower than average numbers of welfare recipients and handicapped were penalized with higher standards. That is, DOL adjusted performance standards for the effect of the training center's enrollment policies on  $P$ .

Because some enrollees benefited more from job training than did others, the training center's choice of enrollees affected the total earnings gains it was able to achieve. The DOL adjustment model by controlling for enrollee characteristics ensured that training centers

would not favor high-h persons over low-h persons. This neutrality might have been counterproductive, however. Low-h persons (persons who had low human capital) may also have been less able to benefit from job training. In terms of the model, low-h persons may have also had low human capital impact, that is, low  $k$  (for evidence on this, see Heckman, Heinrich, and Smith, 2002 and Heckman, Smith, and Clements, 1997). DOL may have been willing to trade away some efficiency for a more equitable allocation of job training resources, however.

While the DOL adjustment method adjusted standards for enrollee characteristics, it did not adjust for the training center's training choices. In the first seven years of JTPA, the performance measures evaluated employment outcomes only once and only on the enrollee's final day of training. In the absence of controls for the services offered, the performance measures created incentives to emphasize short run, "quick fix"-type job placement activities in lieu of longer-term activities with more training content (for evidence of this, see Courty and Marschke, forthcoming, and Marschke, 2003). DOL's move to more long-range measures of employment (the follow-up measures described in Table 1) in the 1990s probably reduced this kind of adverse selection.

## **Dynamics**

The practice of pegging the performance standard to the performance of the training center at the 25<sup>th</sup> percentile in a previous period caused the effort exerted by training centers to change over time and was likely unsustainable. To see this, assume a stationary economy and that training centers are rewarded for meeting or exceeding a standard. Suppose, as in JTPA, a base year's performance is used to set the standard in year 1, the first year in which training centers are subject to the performance-based incentive system. The standard in year 1 is thus the performance of the training center at the 25<sup>th</sup> percentile of base year performance. If training centers respond to incentives, then in year 1, as training centers strive to exceed the performance standard, the distribution of training centers' performances should shift to the right. This rightward shift of performance outcomes means that the new 25<sup>th</sup> percentile, which becomes the basis of the standard in year 2, exceeds the old 25<sup>th</sup> percentile. Thus, the standard training centers confront in year 2 is higher than the standard in year 1. In year 2, to meet the new

standard, training centers must exert more effort than they had in previous periods, shifting the performance distribution to the right yet again, which results in a higher standard for year 3, and so on. Because the standard is growing ever higher, the amount of effort necessary to meet the standard also grows, and thus outcomes grow, which in turn increases the standard. Eventually, however, the standards reach a height that discourages effort, leading to an increase in training center failure rates.

Table 2 reports the departure points for a number of the original JTPA performance measures. Table 2 includes departure points for the standards for the adult employment rate at termination measure, the adult welfare employment rate at termination measure, and the youth employment rate at termination measures. These departure points were consistently set at the 25<sup>th</sup> percentile of a previous year's distribution of outcomes. (The wage and cost standards are excluded from the table as DOL set them based on factors other than the previous year's system-wide performance. See p. 7254 of *Federal Register*, Volume 53, Number 4, March 7, 1988. The table also excludes the standards related to the second generation measures, that is, the follow-up measures described in Table 1.)

For the performance measures represented in Table 2, the 1984 standards were set using data from the final year of JTPA's predecessor job training program (wherein training centers were not subject to performance-based incentives, but did have their performance measured to lay the groundwork for JTPA). The 1985 standard was recalibrated using system-wide performance over the first nine months of the JTPA program (a trial period where no incentives were paid). Starting in 1986, DOL issued new departure points every other year. The departure points for 1986 and 1987 were set at the 25<sup>th</sup> percentile of program year 1984 performance. The departure points for 1988 and 1989 were set at the 25<sup>th</sup> percentile of program year 1986 performance.

Table 2  
Departure Points for First Generation JTPA Standards

Program year	Adult Employment Rate at Termination	Adult Welfare Employment Rate at Termination	Youth Employment Rate at Termination

---

1984	47.0	NA	21.4
1985	57.1	NA	36.4
1986 and 1987	62.4	51.3	43.3
1988 and 1989	68.0	56.0	45.0

---

## Notes:

1. DOL eliminated both adult employment rate at termination measures from its core group of measures at the end of program year 1989.
2. DOL included the youth employment rate at termination beyond 1989. Because 1988 outcome data were unavailable at the beginning of 1990, DOL left the 1990-1991 departure point for this measure unchanged from the previous period.

Table 2 shows a general increase in departure points at the beginning of the JTPA program. Table 2 shows that the departure points in 1986-87 were much higher than those in 1984-85. This is understandable. The departure points for 1984-85 were based on performance under JTPA's predecessor program and performance during the initial nine months of JTPA, during which training centers were not subject to incentive policies. The departure points in 1986 and 1987, however, were based on program year 1984 performance, which occurred when training centers' performances were subject to an award.

Effort was likely subject to diminishing returns so that rising standards became more and more difficult to meet. This suggests that over time the year-to-year increases in departure points would diminish. Note that the increase in departure points between 1986-87 and 1988-89 was smaller than the previous period's increase. The reader should note also, however, that over the period represented by this table, labor markets were tightening. The upward economic trend may have been responsible for some or all of the increase in standards reported in Table 2.

## WIA

In 2000, the Workforce Investment Act of 1998 supplanted JTPA. Heinrich (this issue) presents a detailed assessment of performance management under WIA. (For a description of WIA's performance measurement system, see Barnow and Smith, 2002.) In this section, we briefly discuss the issues that are relevant to performance benchmarking in light of the principles of information economics exposed earlier.

WIA changed the character of the performance measurement system in a number of ways. As in JTPA, under WIA most of the performance measures are based on the labor market outcomes of enrollees. WIA does add several new measures, including subjective measures of enrollee and employer satisfaction. More importantly, for the purposes of this paper, WIA discarded the formulaic approach to constructing standards in favor of one based on negotiation between the states and DOL. In addition, under WIA, state budgets are now contingent upon meeting *state-level* standards. For most states, past performance outcomes are the starting points for this negotiation. While standards in WIA are no longer adjusted formulaically for local economic factors, WIA permits such factors to influence the negotiations. At the start of WIA, in 2000, standards were set for WIA's first three years: 2001, 2002, and 2003. Thus, the standards in 2003 are based on economic predictions made over three years before. While WIA appears to have given states the option to re-negotiate the standards (for example, when the state economies performed worse than expected), few states appear to have taken advantage of this. This has led some policy-analysts and scholars to call for the re-instatement of JTPA-like adjustment models (Barnow and Smith, and Heinrich).

Interestingly, DOL requires states to demonstrate improved performance from year to year. Thus states and DOL have built in yearly increases in the standards. The states then pass on standards that reflect the state-level standards to the individual training centers. As Heinrich notes, this creates a ratchet effect whereby states have an incentive to extract less than the efficient level of effort from their training centers in earlier years.

The presence of conditions that lead to ratchet effects in both JTPA and WIA, against theoretical predictions, suggests that the design of performance incentives may follow a different logic than just an economic one. A possible interpretation is that showing increasing performance over time may be beneficial for political reasons to demonstrate progress and gain public approval.

## **Conclusions**

This paper reviews the economic literature on information economics and draws simple lessons for the construction of performance standards. We demonstrate the relevance of these

lessons in the context of the federal job training programs under JTPA and WIA. We find that the designers of the measurement system have leveled the playing field to provide even performance incentives across the entire organization. The designers have established a system to take into account shocks that are outside the agent's control to reduce the risk faced by the agent. The designers have also tried to reduce the potential negative distortions due to hidden information.

Finally, not all evidence is consistent with the theory. We identify some negative dynamic properties of the incentive system that made it unsustainable in the long run. Our analysis reveals that the dynamics of performance benchmarking in both JTPA and WIA may introduce inefficiencies by giving an incentive to the entire organization to demonstrate low performance early on to allow for progress as the program develops. These dynamic properties violate economic principles and may be explained by a political logic. Understanding the rationale of this dynamic could be a promising line of research.

## **Notes**

(1) Note thus the economic theory of information provides a theoretical framework for the growing literature in public administration (see, e.g., Rubenstein, Schwartz, and Stiefel, 2003, on adjusted performance measures).

(2) This is in contrast to the simple principal-agent model described above. In the model described above, we portrayed the agent as a single worker. Under JTPA, the agent was an agency and the budgetary award was a function of the collective effort of all bureaucrats in the agency. Unlike individual incentives, group incentives are subject to the classic free-riding problem. By increasing her effort, any single caseworker in the training center raises not only her award, but also raises the award for all others. Under group incentives, because she does not enjoy the full benefit of her effort, she may exert too little effort. While we suspect that budget-based awards and the free-riding problem muted JTPA's incentives somewhat, we suspect JTPA's group incentives may have been sufficiently strong to change individual caseworker effort.

(3) Barnow (1992) writes that “[w]hen estimated coefficients have an unexpected sign, the variables are dropped from the models and regressions are re-estimated.” (p. 292) For example, in some regressions, DOL dropped an indicator variable for Hispanic enrollees because it apparently showed a positive effect on performance outcomes.

## **References**

Anderson, Kathryn and Richard Burkhauser and J. Raymond. 1993. “The Effect of Creaming on Placement Rates under the Job Training Partnership Act.” *Industrial and Labor Relation Review*. 46(1): 613-624.

Barnow, Burt. 2000. “Exploring the Relationship between Performance Management and Program Impact.” *Journal of Policy Analysis and Management*. 19(1): pp. 118-141.

Barnow, Burt S. and Jeffrey A. Smith. 2002. What Does the Evidence from Employment and Training Programs Reveal about the Likely Effects of Ticket-to-Work on Service Provider Behavior? Working paper. University of Maryland.

Bouckaert, Geert. 1993. Measurement and Meaningful Management. *Public Performance and Management Review* 17(1): 31-43.

Brooks, Arthur C. 2000. The Use and Misuse of Adjusted Performance Measures. *Journal of Policy Analysis Management* 19(2): 323-29.

Courty, Pascal and Gerald Marschke. 2004 (Forthcoming). Making Government Accountable: Lessons from a Federal Job Training Program. *Public Administration Review*.

Courty, Pascal and Gerald Marschke. 2003. “Performance Funding in Federal Agencies: A Case Study of a Federal Job Training Program.” *Public Budgeting and Finance*. 23(3): pp. 22-48.

Cragg, Michael. 1997. Performance Incentives in the Public Sector: Evidence from the Job Training Partnership Act, *Journal of Law, Economics, and Organization*. 13(1): pp. 147-168.

Dickinson, Katherine P., Richard W. West, Deborah J. Kogan, David A. Drury, Marlene S. Franks, Laura Schlichtmann, and Mary Vencill. 1988. Evaluation of the Effects of JTPA Performance Standards on Clients, Services, and Costs. Research Report No. 88-16, National Commission for Employment Policy.

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

Hatry, Harry. 1999. Mini-Symposium on Intergovernmental Comparative Performance Data. *Public Administration Review* 59(2): 101-104.

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

Heckman, James J. and Jeffrey A. Smith 2003. "The Determinants of Participation in a Social Program: Evidence from the Job Training Partnership Act," IZA Discussion Paper no. 798.

Heckman, James J., Jeffrey A. Smith, and Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies*. 64(4): 487-535.

Heckman, James J., Jeffrey A. Smith and Christopher Taber. 1996. "What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance in the JTPA Program." In Gary Libecap, ed. *Advances in the Study of Entrepreneurship, Innovation, and Growth*, Vo. 7. Greenwich, CT: JAI Press, pp. 191-217.

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

Heinrich, Carolyn. 2003. Improving Public-Sector Performance Management: One Step Forward, Two Steps Back? Working paper. University of Wisconsin.

Johnston, Janet W. 1987. *The Job Training Partnership Act: A Report by the National Commission for Employment Policy*. Washington, D.C.: U.S. Government Printing Office.

Kravchuk, Robert and Ronald Schack. 1996. Designing Effective Performance-Measurement Systems under the Government Performance and Results Act of 1993. *Public Administration Review* 56(4): 348-358.

Marschke, Gerald. 2003. *Performance Incentives and Organizational Behavior: Evidence from a Federal Bureaucracy*. Working paper. University of Albany, State University of New York.

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

Rubenstein, Ross, Amy Ellen Schwartz, and Leanna Stiefel. 2003. Better than Raw: A Guide to Measuring Organizational Performance with Adjusted Performance Measures. *Public Administration Review* 63(5): 607-615.

Stiefel, Leanna, Ross Rubenstein, and Amy Ellen Schwartz. 1999. Using Adjusted Performance Measures for Evaluating Resource Use. *Public Budgeting and Finance* 19(3): 67-87.

Taylor, Frederick Winslow. 1911. *Principles and Methods of Scientific Management*. *Journal of Accountancy* 12(2): 117-24.

Wholey, Joseph. 1999. Performance-Based Management. *Public Performance and Management Review* 22(3): 288-307.

Wholey, Joseph and Harry Hatry. 1992. The Case for Performance Monitoring. *Public Administration Review* 52(6): 604-610.

## Appendix

Although the DOL allowed states some flexibility in developing standards, most states used the DOL's adjustment methodology (see, e.g., Courty and Marschke, 2002), described here. The adjustment methodology worked as follows. Consider an arbitrary performance measure and let  $P_l$  be the outcome produced by training center  $l$ . The DOL adjustment scheme posited that the following function generated performance outcome  $P_l$ .

$$P_l = \mathbf{a} + \mathbf{b}_1(x_{1l} - \bar{x}_1) + \mathbf{b}_2(x_{2l} - \bar{x}_2) + \dots + \mathbf{b}_M(x_{Ml} - \bar{x}_M) + \mathbf{e}_l, \quad (1)$$

where  $x_{1l}, x_{2l}, \dots, x_{Ml}$  are training center  $l$ 's realizations for the  $M$  factors chosen,  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_M$ , are the average realizations of these factors over all JTPA training centers, and  $\mathbf{e}_l$  is a site-specific error term. Biannually DOL estimated the coefficients  $\mathbf{b}$  with the most recent two years of training center-level data using ordinary least squares.  $\mathbf{b}_m$  expressed the impact of an increase in the factor  $x_{ml}$ , on the outcome  $P_l$ , holding other factors constant. DOL chose a different set of factors for each performance measure. DOL chose those economic factors and demographic variables based upon their availability and whether the factors were statistically correlated with the performance outcomes. In addition, political considerations may have played a role. (3)

Table A presents an example of a JTPA worksheet for adjusting the adult employment rate at termination in 1987. The first six adjustment factors in the table are enrollment population characteristics (the percentage of the participant population that is female, black, Hispanic, Asian, handicapped, and welfare recipients). The last two adjustment factors are measures of the local economy (unemployment rate and population density). Column B presents factor values for a hypothetical training center. Columns C and E present the actual national factor averages and the weights from DOL adjustment model for 1987. These weights are the estimated effects of each characteristic on the performance outcome adult employment rate at termination (estimated  $\mathbf{b}$ 's from equation 1). The training center's realization of each of the factors is compared to the national average and the difference is multiplied by the appropriate

weight. For example, suppose the hypothetical training center served 1000 persons during 1987 of which 499 were female. Thus, its percentage female factor was 49.9 percent. To obtain the adjustment to the standard for the female participation factor, one multiplies the difference between the training center's factor value and the national average ( $49.9 - 52.8 = -2.9$ ) by the adjustment weight ( $-.20$ ). The adjustment weight reflected how the enrollment of women historically affected the employment rate outcome (the negative weight implies that on average, compared to men, women produce a lower employment rate at termination).

TABLE A  
Department of Labor's Performance Standard Adjustment Model  
Performance Standard: Adult Employment Rate at Termination

A. Local Factors	B. Training Center Factor Values	C. National Averages	D. Difference (B-C)	E. Weights	F. Effect of Local Factors
% Female	49.9	52.8	-2.9	-.020	.058
% Black	41.2	23.8	17.4	-.081	-1.41
% Hispanic	30.1	7.9	22.2	-.009	-.20
% Asian	2.1	2.4	-.3	-.022	.01
% Handicap'd	9.5	9.1	.4	-.093	-.04
% Welfare Recipient	35.0	29.8	5.2	-.276	-1.44
Unemploy'm't Rate	8.8	8.0	.8	-.623	-.50
Population Density	.21	.6	-.39	.771	-.30
G. Total Effect of Local Factors on Performance Expectations					-5.77
H. National Departure Point					62.40
I. Model Adjusted Performance Level (G+H)					56.60

Notes:

1. Local Factors listed in Column A are given by the Department of Labor. Percentages are of year's participant population.
2. Values for Columns C, E, and H are given by the Department of Labor.
3. Values for Column B are for a hypothetical training center.

Weighted differences for the other factors were calculated similarly. The effects of local factors on performance expectations

(Column F)— were summed [Line G] and added to the national departure point. The departure point (the 25th percentile value) for the measure was 62.4. The final performance standard (56.6) is the sum of the departure point (62.4) and adjustment factor (-5.8). The state used the final standard to establish whether the training center had met its adult employment rate target.

DOL intended that the bar be set to a height appropriate to the training center's circumstances. Thus it included measures of the local unemployment rate and of local population density to capture aspects of the local labor market in which the training center operated. For example, as one can see from Table A the weight on the local unemployment rate measure was negative: a one point increase in the local unemployment rate lowered the standard by about two-thirds of a point. Because training centers were small relative to the local labor market, the unemployment rate is an example of an influence on the performance outcome, which was likely to be beyond the training center's control.

As Table A also makes clear, an important class of characteristics for which DOL adjusted standards was the composition of the enrollment pool. While the enrollment pool reflected in part the composition of the local eligible population (an influence beyond the training center's control), it was at least partly a choice variable. Adjusting the performance standard in this way may have discouraged cream-skimming.