

# **SETTING THE STANDARD IN PERFORMANCE MEASUREMENT SYSTEMS**

**PASCAL COURTY, CAROLYN HEINRICH AND GERALD  
MARSCHKE**

## **Query Sheet**

Q1 Au: 2002 or 1999. pls check?

Q2 Au: Refs. not cited in text. Pls cite in text or delete from ref. list?

## SETTING THE STANDARD IN PERFORMANCE MEASUREMENT SYSTEMS

PASCAL COURTY  
EUROPEAN UNIVERSITY INSTITUTE

CAROLYN HEINRICH  
UNIVERSITY OF WISCONSIN-MADISON

GERALD MARSCHKE  
UNIVERSITY AT ALBANY, SUNY

5

**ABSTRACT:** *A fundamental challenge in the design of performance measurement and incentive systems is the establishment of appropriate benchmark levels of performance, also known as performance standards. Drawing from the information economics, contract theory and public administration literatures, we derive theoretical implications for the construction of performance standards. We then assess alternative methods that are commonly used to construct performance standards and consider their application in performance measurement systems in public programs. We draw out important lessons for the establishment of performance benchmarks and other implications for performance standards system design in public organizations.*

10  
15

---

Performance measurement as a management tool dates back more than a century, as 20  
manifest in the scholarship of Woodrow Wilson (1887) and Frederick Taylor (1911),  
who urged a more “business-like” approach to management in general and public  
administration in particular, including “setting standards” and “exercising controls  
to ensure conformity with standards.” These early efforts to improve executive man-  
agement and organizational performance were focused primarily on measuring individ- 25  
ual employee performance, with minimal attention to the role of performance  
incentives in organizations and the “social character” of these systems, as later  
advanced by Chester Barnard (1938). In recent decades, public sector performance

measurement systems have shifted toward an explicit focus on measuring organizational outcomes and rewarding results that meet or exceed standards. Performance standards are now used to assess the outcomes of a wide range of public sector activities—from teenage birth reduction efforts to defense contract procurement—sometimes with high stakes for achieving (or failing to achieve) performance goals. 30

A key element in the design of these performance measurement systems is the establishment of appropriate benchmark levels (or standards) of performance to guide the evaluation of program outcomes. Performance benchmarks serve as a basis for system incentives, and influence the responses of public managers and staff operating a program. In systems with rewards and sanctions linked to results, performing above or below the standards can have important short-term consequences (e.g. budgetary rewards or revisions, positive or negative recognition), as well as long-term ones (e.g. promotion, structural reorganization). 35 40

In this paper, we review literature in information economics, contract theory (see, for example, Prendergast 1999 or Dixit 2002<sup>1</sup>), and public administration to draw out theoretical implications for the construction of performance standards in public organizations. We then assess alternative methods that are commonly used to construct performance standards and consider the relevance of these lessons for the design of performance measurement systems in public programs. Focusing in particular on performance benchmarking in U.S. workforce training programs, (the Job Training Partnership Act, JTPA, and Workforce Investment Act, WIA, programs), we assess whether the design of performance standards in these programs is consistent with basic principles derived from theory. 45 50

We take for granted in our study the decision to introduce performance measurement, and we try to analyze how this can be done effectively. Clearly, performance assessment serves important functions in public organizations other than promoting program efficiency and effectiveness. Marshall et al. (2000) described three primary functions: accountability for public expenditures, the production of comparative information to inform customer choices in public services, and improvement of professional practice and program management. For example, public managers may use performance information to identify best practices or to communicate to outside constituencies legitimate information about organizational achievements. We acknowledge the possibility that the introduction of performance measurement may transform organizations through channels other than those we discuss in our literature review. We likewise recognize that political and ethical concerns also influence construction of performance standards and use of performance data. We discuss these issues to a greater extent in our case study analysis of performance benchmarking in public training programs. 55 60 65

## THEORY-BASED FRAMEWORK FOR PERFORMANCE BENCHMARKING

We frame our discussion in terms of the principal-agent model, a theoretical framework commonly applied in the economics and public administration literatures. 70

There are critics of this model who argue that it overemphasizes the self-seeking behavior of agents and neglects social interactions and motivators. In his classic study of organizations, Thompson (1967), for example, described the importance of cliques, social controls based on informal norms, and status that influence the performance of organizations. Similarly, stewardship theory emphasizes collective goals and public managers “whose motives are aligned with the objectives of their principals,” or who highly value cooperative behavior even when their interests and those of the principal diverge (Davis, Donaldson, and Schoorman 1997, 21). And there is also a growing literature on public-service motivation that asserts individuals act in ways that contribute to the public good, not in response to incentives offered by organizations for performance or commitment, but as a way of satisfying their personal needs (Perry and Porter 1982; Rainey 1982; Wise 2004). Wise (2004, 675) adds that public-service motives have the potential to empower public servants to “overcome self-serving interests, moral inertia and risk avoidance,” and Crewson (1997) links them empirically to greater organizational commitment and lower employee turnover. Although we acknowledge the roles of social and cultural norms, of public-service motivation, and the influence of political or personal power relationships as described in these alternative theoretical frameworks, we rely primarily on principal-agent theory in modeling behavior and relationships in this study.

In our application of principal-agent theory to the study of performance standards systems, we call the party who designs the measurement system the principal and the party whose performance is measured the agent. We denote the measured performance  $P$ , and the benchmark level of performance, or the performance standard,  $P_0$ . The difference between the performance outcome ( $P$ ) and the performance standard ( $P_0$ ) is denoted  $\Delta P$ , that is,  $\Delta P = P - P_0$ .

We are interested in the methods that are used to construct the performance benchmark  $P_0$  and in the kind of information that these methods incorporate in the benchmark. Although we recognize that there may not exist a single method of construction that could be effectively applied in all situations, we still believe that some methods are largely more effective than others. We say that a standard is poorly constructed or “ineffective” if it is missing key pieces of information, and/or if it is likely to send the wrong signals and to stimulate behavioral responses with negative implications. In addition, we recognize that, in practice, organizations will often use multi-dimensional measurement systems with multiple measures and performance benchmarks. Although we explicitly discuss these issues, for the sake of conciseness, we focus in our literature review on the simplest case with a single performance measure, as this is sufficient to highlight the main lessons from the literature without loss of generality.

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  $e$ . We model effort as a one-dimensional choice by the agent. It is useful, however, to think of  $e$  as a vector of activities; the agent chooses not only how hard to work, but also how to allocate her time and effort across different activities.

For example, in job training centers, case workers allocate their effort toward recruiting participants, assessing their training “needs,” networking with other social service organizations, managing contracts with external vendors, and bookkeeping, to name a few of their activities. The lessons we draw from our model based on a simple formulation—assuming that  $e$  is a scalar variable—are robust to this alternative and more realistic assumption. 120

In the simplest formulation, performance is equal to effort  $P = e$ , and value added is equal to  $\Delta P = e - P_0$ . We think of value added here as the agent’s contribution to the principal’s welfare, net of costs. In the case of a job training program, assuming its objective is to raise the earnings and employability of the poor, value added is the value of the labor market skills enrollees acquire due to the exertions of training center workers, net of training costs. 125

As discussed above, setting the absolute level of the performance standard is a critical task in performance measurement systems. In the federal programs we study—JTPA and WIA—and in other similar programs in effect since the 1993 passage of the U.S. Government Performance and Results Act (GPRA), government officials are required to undertake this task annually. We assume (consistent with practice in these programs) that 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 boosts incentives to improve performance, because the agent has to supply more effort to meet the standard and avoid sanctions. 130 135

Large or incessant increases in the standard, however, also diminish the credibility of the measurement system, with the consequence that the agent may simply give up or search for alternative, possibly unproductive ways to increase measured performance. In our model, 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  $e_0$ . In an efficiently functioning system, 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. In this case, the performance standard is set at the level of performance that occurs when the agent provides the competitive level of effort,  $P_0 = P(e_0)$ . 140 145

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. Agents would not apply or compete for the job or contract. On the other hand, if the performance standard were set below  $P(e_0)$ , the principal would be over-compensating the agent. 150

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 was first used in a systematic way (and is still common practice). Taylor (1911) was perhaps its most famous early proponent. Miller (1992, 102) also used a piece-rate production setting to analyze “managerial dilemmas” and to consider how an incentive system could “harness individual self-interest in pursuit of organizational goals,” “transforming an organizational social dilemma into an organizational ‘invisible hand.’” This choice 160

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 those arising in production manufacturing.

Suppose a manual worker in a factory is paid a wage ( $w$ ) per hour. The wage is paid independently of the level of worker performance. In addition to the fixed wage, 165 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 an hourly rate,  $P_0 = e_0(w)$ , which a represen- 170 tative worker earning  $w$  would achieve, and then use this benchmark to evaluate actual performance. In other words, the principal actually assesses the level of performance occurring 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 interpret- 175 ation, 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 180 excess effort relative to the competitive level of effort. Counterfactual experiments such as time-and-motion studies, however, are practical only in a few occupations, typically involving manual work. Many public and private sector work situations involve non-manual work, complex group interactions, and non-standardized outputs, making experimental studies to construct counterfactual performance 185 benchmarks very costly.

These complications mean that the methods typically used to construct performance standards are imperfect. Real world methods necessarily balance the cost of establishing fair and appropriate standards and the expected return to the organiza- 190 tion from assessing value-added more precisely. Also, of course, in the "real world," non-economic factors—e.g. political goals, legislative requirements, etc.—may influence standard setting processes as well. These challenges and tradeoffs will become evident in the case study analysis of the JTPA and WIA performance standards that follows. First, however, we review a set of generic problems that any method of 195 setting performance targets must address.

### Leveling the Playing Field

Thus 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 differ- 200 ent work conditions. To illustrate, we return to the time-and-motion study example presented above and assume that there are multiple workers assigned to different machines. We also assume that the machines vary in their productivity in the sense

that the amount of effort required to produce a unit of output varies by machine. (One might model this idea formally by defining the performance outcome from machine  $k$  when effort is  $e$  as  $P(e) = ke$ , where  $k > 0$  and  $k$  is different from machine 205 to machine.) Assume that each machine's productivity is known to both the principal and the agent.

If the agent is allowed to refuse to work on a machine, 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 stan- 210 dard for all machines, say  $P_0 = e_0$ , then the agent will only agree to work with machines that exhibit high marginal productivities. An important point is that time-and-motion studies would 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. 215

Consider, for example, the job training caseworker with the responsibility to assess clients and place them into jobs at a rate required by the performance standard. The caseworker will prefer to work with the most motivated and capable clients and direct them into the most effective employment preparation activities. Consequently, 220 in the absence of adjustments to the standard, the caseworker would respond by discriminating against low-ability enrollees and directing only the higher-ability trainees into services with the highest measured performance outcomes (e.g., job placement activities).

### Insurance and Uncontrollable Risk

The time-and-motion study example assumed no uncertainty about the worker's 225 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 external shock or influence, for example, a power outage that slows production (or in the job training example, an economic recession that dampens job placement success). The performance outcome is now equal to  $P + \varepsilon$  and the worker's value added is 230  $\Delta P = (P + \varepsilon) - P_0$ , where  $\varepsilon$  is a mean zero random variable, that is realized only after the agent has chosen his or her level of effort. Setting the standard at a level that does *not* take into account these circumstances or context implies that a worker who supplies the effort level required to achieve the performance standard will sometimes over-perform and other times under-perform relative to the standard. In this 235 situation, 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 performance. Outside shocks, however, do change the realized level of performance and value added, and therefore, the worker's compensation.

Although a risk-neutral worker will not suffer any disutility from this variation in 240 compensation, a risk-averse worker will, and this establishes a first rationale to construct a performance standard that takes such outside influences into account. The merit of performance standards will then depend in part on their ability to control for outside risk (i.e., circumstances beyond the control of public managers or staff). The logic of agency theory is that standards that properly account for external 245

influences reduce compensation risk, thereby increasing the agents' welfare. Lowering the risk faced by the agent is also desirable for the principal, as it means that he does not have to offer the agent a higher wage to compensate for risk bearing.

The key challenge for performance standards system designers is to identify the sources of controllable versus uncontrollable factors that influence the performance outcome. In other words, it is important to the extent feasible to hold the worker responsible for effort,  $e$ , but not for external influences,  $\varepsilon$ . At the same time, 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 that the workers' machine breaks down if the agent might have anticipated and prevented the breakdown. In the context of the job training example, the principal *does* want to hold the case worker responsible (and reward) efforts made to appropriately assess clients and facilitate better worker–employer matches. But it would be unfair and inefficient to penalize the case worker (or job training center) for a lower rate of worker/employer matches if it is due to a declining number of job opportunities.

As suggested above, though, risk aversion is not the only reason the agent may experience disutility from performance standards that fail to control for outside risks. Another closely related concern is fairness. The issue of income variability drives the concern under risk aversion, while other considerations, such as interpersonal comparisons, may foster concerns about fairness. 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, compared to a group shock that lowers all workers' performance. The former shock generates different treatments among individuals who have essentially behaved identically. As where there is risk aversion, if workers value fairness, 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 setting performance standards, the principal may want to take into account information about shocks that influence the performance outcome and are outside the agent's control. These shocks are observed only after the agent has chosen a level of 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 her level of effort. For example, assume that when the agent is assigned to a new machine, she alone knows the productivity of that machine and can use this information to make her effort choice. We say that the information is privately known by the agent, and consistent with the literature, we characterize such situations as hidden information (Holmstrom 1982; Miller 1992).

Hidden information further complicates the problem of setting standards. To illustrate, we return to the job-training program example and again assume that the training efforts of case workers in this program are evaluated in part according

to the rate at which their clients secure jobs. Case workers observe relevant information about applicants' likely success in training activities (e.g., personal 290 motivation and employment barriers), and based on this information, they 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 performance measure are not necessarily those who benefit most from training. Indeed, some applicants to the job training program may be highly likely to obtain employ- 295 ment on their own. As a result case workers 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 300 participant is enrolled. We denote hidden information by  $h$  to distinguish this kind of information from information that is publicly known, such as the information about varying productivities of machines. If the caseworker observes the applicants' type, she has an incentive to enroll only easy-to-serve applicants because they produce better outcomes. 305

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, adjusting upward the performance of those agents who have enrolled a larger fraction of hard-to-serve applicants. By assumption, however, only the agent knows this information. If the principal were to ask the agent what type of participants she has enrolled, the agent 310 would have an incentive to report enrolling only hard-to-serve enrollees, and the principal would have no way to verify that the agent is telling the truth.

Practically speaking, the principal could correct these distortions by developing a specific measure to target the hard-to-serve groups, for example, using observable variables such as welfare receipt or limited English proficiency as proxies for 315 "hard-to-serve." This would assume that welfare recipients or non-native English speakers are harder to serve because they require higher investments for equal outcomes. The principal could set a lower performance standard for these individuals, for example,  $P'_0 = e_0 - h$ , where  $P'_0 < P_0$  by construction. Of course, we know that not all welfare recipients (or non-native English speakers) are identical; some are 320 easier to serve than others, and the agent observes this information. Again, the agent will be inclined to select a non-representative sample of these groups. This implies that the principal has corrected some distortions because the agent's attention is now focused on a needier target population, but the agent will still select those applicants who are the easiest to serve within these sub-populations of applicants. 325

Note that a slightly different problem from hidden information, known as adverse selection in the 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, observed privately, of meeting the standard. In the job training case, this happens when there are multiple case workers who face different eligible populations 330 and when the case workers privately observe this information. The distinction between hidden information and adverse selection 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, 335 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 340 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. However, these agents may not be those who generate the highest returns from the principal's perspective. 345 The agent's selection rule poses a problem when it does not correspond to the rule that the principal would use, had the principal had the same information as the agent.

### Multiple Principals

Another distinctive feature of performance standards systems in the public sector 350 is the greater likelihood that agents will work for more than one principal. In the context of public organizations, one should think of principals as a widely defined category that includes all constituencies or interest groups that may influence the actions of the agent, either directly through explicit rewards or indirectly through more subtle channels. For example, in the context of our application to the JTPA 355 and WIA programs, Congress, the U.S. Department of Labor, and state governments would be the main principals, since these are the actors who directly define the goals and activities of the organization, through the design of the incentive system and performance standards, and also through other organizational features. But local politicians, private industry council representatives, and other interest groups 360 should also be viewed as secondary principals, since these parties likewise have roles in influencing training program priorities and agency actions.

The key implication of the presence of multiple principals is increased complexity in the choice of performance measures, particularly if the interests of the different principals are not aligned, that is, they emphasize different priorities or outcomes. 365 The agent has to choose how to allocate her effort level across the various goals or objectives of the principals, which might be represented in a performance standards system by multiple standards,  $P_1$ ,  $P_2$ ,  $P_3$ , etc.

Q1 Dixit (1999) proposes an analysis of multiple principals competing non-cooperatively for the agents' effort. As expected, the agent will allocate more effort toward 370 the objectives of principals who compensate (with wage,  $w$ ) at a higher rate (or provide greater rewards for achievement in some form or another). In other words, if  $w_1 > w_2$ , then  $e_1 > e_2$  and  $P_1 = e_1(w_1) > P_2 = e_2(w_2)$ ; performance is higher on the outcome set by the principal who calls for  $P_1$  and provides greater rewards for its achievement. Dixit demonstrates that the marginal level of effort applied by the 375 agent ( $e_1$ ,  $e_2$ ,  $e_3$ , etc.) toward the achievement of the various outcomes will decrease

with the number of principals. The reason is simply that each principal will reward the agent for success on the particular dimension(s) of effort that concern him or her, but she will also insure the agent against failure on dimensions of effort that concern the other principals. If principals choose the level of incentive non-cooperatively, the 380 desire to insure the agent will conflict with the desire to provide incentives.

In investigating how the principals compete for the agent's effort, Dixit also shows that the declines in agents' marginal level of effort (as the number of principals increases) will be exacerbated if the efforts across principals' objectives are substitutes. In other words, the principals undermine one another, and the impact of 385 the incentives is diminished. In equilibrium, all principals call for effort, but since efforts are substitutable, the incentive effects on total effort are reduced.

Dixit's analysis suggests two recommendations for organizational design. First, one should allocate and organize tasks across agents based on whether they are complements or substitutes. Complementary activities can be grouped together, but the 390 grouping of substitute activities should be avoided. In the context of the JTPA program, if there are some principals who are more concerned about equity of allocation (local government) and others more concerned about efficiency (the federal government), it may be optimal to divide up the functions of enrollment and training, and to assign each of these activities to two separate agencies. 395

In addition, the model has implications for how the principals should be allowed to compete. In particular, the principal  $i$  should not be permitted to excuse or cover for the agent's poor performance toward meeting principal  $j$ 's objective. This "compartmentalization principle" has implications in a public organization. Consider, for example, the conflict between enrollment and training in the JTPA 400 program described above, and assume that the proposed solution of breaking up these tasks is not feasible for administrative or practical reasons. In this situation, the principals who are concerned primarily with reaching hard-to-serve populations will try to set the performance standard in such a way that training agencies are not penalized for achieving low performance outcomes. Similarly, principals who 405 care mainly about efficiency will try to minimize the emphasis placed on enrollment choices. A possible result would be that agencies would face low performance standards and no constraints on enrollment. To avoid this outcome, one would want to minimize principals' interference with one another in the setting of performance standards. 410

Although there is considerable discussion of multiple principals in the literature, there is little mention of situations in which there may be a hierarchy among the principals. The political science literature discusses "political multidimensionality" and the difficulty of identifying an "ultimate principal," e.g., the competing interests of House and Senate chambers, committees and other political actors that have 415 implications for the stability of agents' behavior (Maltzman and Smith 1994). However, it is also possible that in a political hierarchy such as that established in the JTPA system, with service providers taking signals from local job training authorities and state/federal policy directives at the same time, agents might allocate their effort toward alternative objectives of these principals according to the principals' 420 position in this hierarchy.

### 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 she acquires new information about the effectiveness of different measurement schemes or about the influence of external factors on performance. 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. In both the JTPA and WIA programs, this has been a central component of the performance standard setting process. The WIA legislation explicitly identifies “continuous performance improvement,” in which performance targets increase each year, as a central tenet of the performance standards system. Such a rule also implicitly exists in any organization that uses past agent performance to set standards for the present. Assume the agent systematically outperforms the standard, and the principal consequently increases it. It could be 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 (and the level of effort required to obtain the same reward in the future). Thus, a simple static view of incentive systems may fail to capture such behavioral responses that arise only when one considers the dynamic nature of performance measurement.

In the economics and management literature, this phenomenon is known as the ratchet effect (Holmstrom and Milgrom 1987; Miller 1992). The agent’s belief about the principal’s policies regarding future standards will significantly influence 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. Miller (1992, 157) likewise recognized the importance of trust in these situations, noting that “‘trustworthiness’ on the part of managers seems to be a necessary element of an effective incentive system.” Another way the principal can eliminate the ratchet effect is by committing never to change a standard, or more realistically, by committing to strict rules for any changes. Such commitment is likely to eliminate fear of the ratchet effect and reinforce incentives for effort.

### OVERVIEW OF PERFORMANCE STANDARD SETTING APPROACHES

We now present a brief overview of alternative methods for constructing performance standards (informed by the theoretical discussion above) and consider the environments where these methods are likely to work well.

### Estimating the Production Function

Most basically, the public agency can attempt to estimate the production function (i.e. the level of productivity expected from a given level of effort) to set the standard. It is sometimes possible to establish a standard through experimentation or through statistical methods. Such an approach will only be valid, however, for production processes that are stable over time and across environments. This is relatively rare, for example, in public social service provision. The use of data on past performance outcomes to construct estimates of the production function is a more common application of this method. A potential problem with this method, as discussed above, is the introduction of a ratchet effect if higher performance outcomes increase future standards. This method is also unlikely to work well in non-stationary environments where the production technology is subject to transient shocks.

### Relative Performance Evaluation

Relative performance evaluation (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 (such as in the Job Corp Center annual performance rankings or "league tables" used for education or health care in the United Kingdom). Alternatively, the principal could compare 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 (Hatry 1999). Of course, this method has its limitations, too, in that it may exacerbate competition and may also result in wasteful behaviors (e.g., sabotage, monitoring others, etc.).

### Negotiating the Standard

With this method, the principal and agent agree on the performance standard. If objective information on the production function is absent and relative performance evaluation is not a viable alternative, this may be the only solution available. This approach requires an environment of mutual trust between the agent and principal(s), that is, one in which the agent does not withhold important information about her effort and capabilities, and where the principal can be trusted to use performance information fairly, for example, not to increase the standard in the event of performance outcomes above the standard. The resulting performance standard (and the corresponding distribution of risk between the principal and agent) may be more a function of the relative bargaining ability of the parties, however, rather than reflecting principles of effective performance standard-setting processes.

## PERFORMANCE STANDARD SETTING IN FEDERAL JOB TRAINING PROGRAMS

In the U.S. government's largest job training program, a performance measurement system has existed for over two decades. During this time, individual training providers have been evaluated by their performance relative to specific, numerical standards. Congress has also legislated important changes in the formulation of these numerical standards. A major redesign of the program five years ago introduced an entirely different approach to setting performance standards, and we will devote considerable attention to the implications of these changes for the system's incentives and functioning. 505

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. In 1982 the Job Training Partnership Act transformed the federal agency administering 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 whom 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 performance standards. 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, Courty and Marschke 2003; Dickinson et al. 1988; Johnston 1987.) 510 520 525

Under JTPA, Congress, the U.S. Department of Labor (DOL), and state authorities shared in designing and implementing the program's incentive policies. The DOL established expected performance levels using a regression-based model with national departure points. States could use the optional DOL adjustment model or develop their own adjustment procedures, although the state-developed procedures and any adjustments made by the governor had to conform to the DOL's parameters (Social Policy Research Associates 1999). A majority of states adopted these models and used the DOL-provided performance standards worksheets to determine performance targets, although some with modifications. 530

The WIA program that replaced JTPA in 2000 introduced a new approach to setting performance standards that involves the negotiation of performance targets. States negotiate with the DOL and local workforce investment areas to establish performance standards, using estimates based on historical data (or past performance) that are intended to take into account differences in economic conditions, participant characteristics and services delivered. The pretext for making this change to a system of negotiated standards was to promote "shared accountability," described as one of the "guiding principles" of the Act (U.S. Department of Labor, Employment and Training Administration 2001, 8). 535 540

In our case analysis of the JTPA and WIA performance measurement systems, the DOL, Congress, and the states constitute multiple principals in the organizational structure, while local implementing authorities (government entities or 545

training centers) function as the agents, undertaking the business of enrolling, training, and finding employment for program clients. Table 1 shows the performance measures currently in effect in the WIA program and also indicates which of these are new to WIA (i.e., that were not used in the JTPA program). Many of these 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 “counterfactual” level of performance, i.e., the level of performance that would occur under a competitive level of effort. Consider, for example, the entered employment (or job placement) rate measure. What employment rate outcome would an agent who supplies a competitive level of effort achieve? Assuming, as discussed above, that the DOL uses past performance in a representative training environment to determine the expected distribution of performance, then differences in employment rate performance outcomes should only be due to differences in effort. For example, if the DOL believes that only 50 percent of the training centers have supplied at least the competitive level of effort, the performance standard should be set at the 50th 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 25th percentile of performance among all training centers, implying that 25 percent of the training centers in the previous program were not supplying a competitive level of effort.

Under WIA, this more systematic approach for setting standards was abandoned. With discretion for setting performance standards transferred to the negotiation process between states and localities, use of past performance information varied widely. The DOL did provide some guidance for negotiated targets under WIA, using data on the performance of seven early implementing states. However, among the majority of states that used baseline performance measures in determining appropriate levels for the standards, the sources for these data differed considerably. The various types of data used included: the projected national averages for the negotiated standards provided by the DOL; federal baseline numbers (available in the federal performance tracking system, i.e., Standardized Program Information Reporting [SPIR] data); unemployment insurance data, and states' own performance baselines from previous program years. Georgia, for example, used program year (PY) 1998 state performance records combined with the projected national averages in negotiations with regional office representatives and local-level officials to determine the performance targets for the first three years of WIA. Some states, such as New Hampshire and Ohio, used unemployment insurance data from earlier periods (PY 1994–1997) combined with DOL performance data available in the SPIR to set performance levels. These considerable differences across states in the performance standard setting process have important implications for the ability of the principal to create a level playing field for all agents.

**TABLE 1**  
Performance Measures—JTPA and WIA

<i>Performance Measure</i>	<i>Description</i>
<b>Adults</b>	
Entered employment rate	The percentage of adults who obtained a job by the end of the first quarter after program exit (excluding participants employed at registration).
Employment retention rate at 6 months	Of adults who had a job in the first quarter after exit, percent age with a job in the third quarter after exit.
Average earnings change in 6 months	Of those who had a job in the first quarter after exit, the post-program earnings increases relative to pre-program earnings.
Employment and credential rate*	Of those adults who received WIA training services, the percentage who were employed in the first quarter after exit and received a credential by the end of the third quarter after exit.
<b>Dislocated workers</b>	
Entered employment rate	The percentage of dislocated workers who obtained a job by the end of the first quarter after program exit (excluding those employed at registration).
Employment retention rate at 6 months	Of those who had a job in the first quarter after exit, the percentage of dislocated workers who have a job in the third quarter after exit.
Earnings replacement rate in 6 months	Of those who had a job in the first quarter after exit, the percentage of pre-program earnings that are earned post-program.
Employment and credential rate*	Of those dislocated workers who received WIA training services, the percentage who were employed in the first quarter after exit and received a credential by the end of the third quarter after exit.
<b>Older youth (19–21)</b>	
Entered employment rate	The percentage of older youth who were not enrolled in post-secondary education or advanced training in the first quarter after program exit and obtained a job by the end of the first quarter after exit (excluding those employed at registration).
Employment retention rate at 6 months	Of those who had a job in the first quarter after exit and were not enrolled in post-secondary education or advanced training in the third quarter after program exit, the percentage of older youth who have a job in the third quarter after exit.
Average earnings change in 6 months	Of those who had a job in the first quarter after exit and were not enrolled in post-secondary education or advanced training, the post-program earnings increases relative to pre-program earnings.

(Continued)

**TABLE 1**  
Continued

<i>Performance Measure</i>	<i>Description</i>
Older Youth Employment/ education/ training and credential rate*	The percentage of older youth who are in employment, post-secondary education, or advanced training in the first quarter after exit and received a credential by the end of the third quarter after exit.
Younger youth	
Retention rate	In employment, post-secondary education, advanced training, apprenticeships in third quarter after exit
Skill attainment rate	Attain at least two goals relating to basic skills, work readiness, skill attainment, entered employment and skill training
Diploma rate	Earn a secondary school diploma or its recognized equivalent (GED)
Customer satisfaction	
Participant satisfaction*	The average of three statewide survey questions, rated 1 to 10 (1 = very dissatisfied to 10 = very satisfied), asking if participants were satisfied with services, if services met customer expectations, and how the services compared to the "ideal set" of services
Employer satisfaction*	The average of three statewide survey questions, rated 1 to 10 (1 = very dissatisfied to 10 = very satisfied), asking if employers were satisfied with services, if services met customer expectations, and how the services compared to the "ideal set" of services

\*indicates measure new to WIA.

### Is the Playing Field Level?

590

Although our discussion thus far has centered on a representative training center, there are important differences across centers in the populations from which they draw their enrollees and in their labor markets. For example, training centers located in relatively depressed labor markets should reasonably expect lower performance outcomes than those located in relatively tight labor markets. The DOL recognized this problem and provided states with a method to adjust standards that took into account features of the training center's population and environment that may have been correlated with the performance outcome. 595

Under JTPA, the 25th percentile (of the distribution of past performance) was established only as a starting or departure point; it was subsequently adjusted using a regression model to take into account the extent to which a training center's characteristics differed from the average center's characteristics. Thus, continuing the above example, the adjustment approach would lower the entered employment rate standard for training centers in depressed job markets relative to those in robust ones. 600

In the WIA program, the formal performance standards adjustment models were discarded by nearly all of the states (the exceptions were Texas, Maryland, and the District of Columbia). At the same time, the DOL instructed states to take into account differences in economic conditions, participant characteristics, and services provided. For a majority, these adjustments to standards were made informally during the review of past performance data and in negotiations. For example, Wisconsin reported using PY 1997 data and the projected averages in negotiations with local officials to set the standards. A comparison of these data shows that when Wisconsin's PY 1997 baseline was above the projected national averages, the projected averages were established as the targets. When Wisconsin's baseline numbers were below the projected national averages, the baseline values were typically set as the targets. Other states, such as Washington, Nebraska, South Carolina and others, followed a similar process.

### Adjusting for Uncontrollable Risks

In addition to accounting for factors (demographic, economic, or other) known at the time performance standards are established, it is important to allow for adjustments to standards that will offset future or unknown risks of poor performance due to conditions or circumstances beyond the control of agents. In other words, the adjustment methodology should also correct for the risk generated by a random shock ( $\varepsilon$ ) in the model. While exceptional performance is still an unbiased estimator of excess effort even in the presence of a random shock, such a shock introduces noise in the measure of value added, and therefore, in the training center's award. Because of risk-averse training staff and uncertain budgets, it is in the principal's interest to formulate performance standards that control both for persistent differences across training centers and for transitory or idiosyncratic shocks.

As described above, many states used past performance data to set performance standards for the first year of the WIA program. In addition, most states also built in anticipated performance improvements for the two subsequent years. However, economic conditions changed significantly between the pre-WIA period and first three years of the program's implementation. Between 1998 and 1999, unemployment rates were declining on average, with 75 percent of all states experiencing a decline. Then between 2000 and 2001, this trend reversed. More than 75 percent of the states experienced an increase in unemployment rates, and the increases were even greater between 2001 and 2002. Table 2 shows that as unemployment rates were increasing and creating adverse labor market conditions for trainees in the first three years of WIA, the standards for performance achievement in the program were increasing.

Year-to-year variations in job availability typically cannot be anticipated by training centers. By adjusting a training center's standards for the local unemployment rate each year, the variance in performance due to unpredictable changes in the environment is reduced. And although these types of adjustments were made in the JTPA system, they were not standard practice under WIA. A 2002 Government Accounting Office (GAO) report confirmed that WIA program administrators were seriously concerned about their ability to meet performance targets. All state

**TABLE 2**  
State Performance and Economic Conditions in the First Three Years of WIA

<i>Performance Goal and Labor Market Conditions</i>	<i>Program Year 2000</i>	<i>Program Year 2001</i>	<i>Program Year 2002</i>
Mean entered employment rate standard for adults	66.44	69.17	70.94
Mean unemployment rate	3.94	4.59	5.35

program administrators reported that some of the performance targets were set too high for them and that the performance standards negotiation processes did not allow for adequate adjustments to varying economic conditions and participant demographics. In fact, the proportion of states meeting or exceeding their performance standards dropped between PY 2001 and PY 2002 for nearly all measures, some dramatically, such as the 21 percent decrease in the proportion of states meeting their older youth entered employment rates (see Table 3).

### Cream-Skimming and Quick Fixes

The pressures generated by a high-stakes performance measurement system can lead to undesirable behavioral responses on the part of agents. The performance standards under both JTPA and WIA were not only “noisy,” but they were also vulnerable to manipulation by agents. One way to increase  $P - P_0$  (and the corresponding performance 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 without receiving training. This behavior has been called cream-skimming (Anderson, Burkhauser, and Raymond 1993; Cragg 1997; Heckman and Smith 2003; Heckman, Smith, and Taber 1996).

To prevent cream-skimming, the DOL adjusted the JTPA standards for the effects of the characteristics of enrollees on  $P$ . As illustrated in the appendix, 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 required to achieve higher standards. That is, the DOL adjusted performance standards for the effect of the training center’s enrollment policies on  $P$ .

These adjustments under JTPA, which apparently did not fully account for all low-h characteristics, may have reduced cream-skimming behavior, but they did not eliminate it (Heckman, Heinrich, and Smith 2002). In addition, the adjustment method did not account for training centers’ choices about the training services made available. Thus, the performance measures generated incentives to emphasize short-run, “quick fix”-type job placement activities in lieu of longer-term activities with more training content (Courty and Marschke 2003). Courty and Marschke (1997, 2004) also showed how program managers strategically managed their “trainee

**TABLE 3**  
 Percent of States Meeting or Exceeding their Negotiated Performance Standards in Program  
 Years 2000–2002

<i>Performance Measure/Standard</i>	<i>Percent of States Meeting or Exceeding Their Negotiated Performance Target</i>		
	<i>Program Year 2000</i>	<i>Program Year 2001</i>	<i>Program Year 2002</i>
Adult entered employment rate	56.7%	66.5%	61.5%
Adult employment retention rate	54.0	60.7	57.7
Adult earnings change	49.3	64.6	48.1
Adult credential rate	36.7	45.6	46.2
Dislocated worker entered employment rate	52.7	65.5	55.8
Dislocated worker employment retention rate	42.0	58.7	51.9
Dislocated worker earning replacement rate	54.7	74.8	61.5
Dislocated worker credential rate	36.7	58.7	55.8
Older youth entered employment rate	58.7	63.6	42.3
Older youth employment retention rate	52.0	61.2	48.1
Older youth earnings change	52.7	64.6	59.6
Older youth credential rate	29.3	31.6	23.1
Younger youth retention rate	38.0	59.2	57.7
Younger youth skill attainment rate	72.0	69.4	53.9
Younger youth diploma rate	25.3	45.6	50.0
Employer satisfaction	45.3	75.7	69.2
Participant satisfaction	51.3	78.6	76.9

inventories” and timed participant program exits to maximize end of the year performance levels.

In the later years of JTPA and under WIA, the DOL made additional changes to the performance standards to reduce incentives for cream-skimming, including a longer follow-up period for outcome measurement and a shift toward measuring changes in outcomes from the pre-program to post-program periods. For example, WIA now measures participants’ employment retention at 6 months (rather than 90 days) post-program, and earnings *changes* and earnings replacement rates at six months are now used instead of earnings *levels* after 90 days.

**Implications of Multiple Principals**

The federal government’s efforts to encourage service delivery to the hard-to-serve and the provision of more intensive training activities were also frustrated by the presence of multiple principals with differing priorities. Although state authorities followed suit in placing more emphasis on these same goals, some local job training authorities continued to demand low-cost placements from their service providers (Heinrich 1999). Heinrich found that service providers were aware of the new federal and state policy directives but focused primarily on job placement rates and cost per

placement in their efforts, largely because these were the outcomes directly rewarded with contract renewals and other forms of recognition at the local level.

The change under WIA to a system in which regional DOL representatives, state authorities, and local representatives engage in negotiations to determine performance standards might have presented an opportunity for greater coordination in aligning these principals' interests and reducing problems associated with divided agent efforts. In practice, however, the lack of formal adjustment mechanisms for standards under the new system only exacerbated these problems. After interviewing WIA program administrators in 50 states and visiting five sites, GAO (2002, 14) concluded that "the need to meet performance levels may be the driving factor in deciding who receives WIA-funded services at the local level." The GAO report and a subsequent study (Heinrich 2004) described how some local areas have limited access to services for individuals who they perceive are less likely to get and retain a job. For example, some have responded to these pressures by augmenting the screening process for determining registrations or by limiting registrations of harder-to-serve job seekers, including dislocated workers whose pre-program earnings were more difficult to replace. A Texas official indicated that even with Texas' relatively sophisticated statistical model for setting and adjusting performance standards, adequate adjustments had not been made for economic conditions.

In her empirical analysis of WIA program performance across the states, Heinrich (2004) estimated OLS regressions using as dependent variables states' actual performance levels, and in separate regressions, the differentials between their actual performance and the negotiated standards. The objective of these analyses was to assess the relationship of local participant characteristics and economic conditions to measured performance and to determine if "adjustments" made in the negotiation process (to establish fair standards) were effective in accounting for these factors. For example, states with a comparatively high number of high school dropouts participating in their programs could have negotiated a lower employment retention rate or earnings change standard in anticipation that their less educated populations would have fewer or less attractive employment opportunities. If the states' initial processes for adjusting performance standards through such negotiations had worked as intended, one would expect to see fewer or weaker relationships between the performance *differentials* and these baseline characteristics (compared to their relationships with actual performance levels). In other words, only state and local program efforts—not characteristics of their populations or economic conditions that were beyond program managers' control—should explain why they met, exceeded or fell below their negotiated performance standards.

Heinrich estimated separate regressions for each of the 17 performance standards for these two dependent variables. In *both* sets of models, characteristics such as race, education, and work history were statistically significant predictors of performance relative to some standards, suggesting that adjustments for participant characteristics were inadequate. In fact, the most consistent, negative predictors of performance levels and differentials were unemployment rates. These findings confirmed that states were not prepared to adjust for what turned out to be significant risks of failure to meet performance standards due to the economic downturn.

**Dynamics**

Under JTPA, the practice of pegging the performance standard to the performance of the training center at the 25th percentile in the prior period likely contributed to unsustainable changes in the level of effort exerted by training centers over time. If training centers responded to incentives and strived to exceed the performance standard, the distribution of training centers' performances would shift to the right, implying that new 25th percentile (which would become the basis of the standard in the next year) would exceed the old 25th percentile. As long as training centers can keep up, the standard grows ever higher, and the amount of effort necessary to meet the standard also increases, leading to higher outcomes and future increases in the standard. More realistically, if performance improvement goes on for long enough, such a system implies that some training centers will eventually fall behind and fail to meet the standard. If such a system were used for long enough, performance improvement should eventually stop and about 25 percent of training centers would perform below the standard.

Table 4 reports the departure points for a number of the original JTPA performance measures (adult employment rate at termination, adult welfare employment rate at termination, and youth employment rate at termination). These departure points were consistently set at the 25th percentile of a previous year's distribution of outcomes. As predicted, Table 4 shows a general increase in departure points over this period of the early JTPA years. The departure points in 1986-87 were much higher than those in 1984-85, which is not unexpected given that they were based on performance under JTPA's predecessor program and the initial nine months of JTPA (during which training centers were not subject to incentive policies).

Under WIA, the DOL strongly encouraged states and localities to set standards that would motivate improved performance from year to year. In fact, in the effort to promote "continuous performance improvement," the states set standards that not only required that they improve over time, but also that the magnitude of the improvements increase from year to year. This approach gave states an implicit incentive to negotiate lower standards in the early years, and some states in fact attempted to do this. North Carolina, for example, was asked by the DOL to increase the level of its negotiated standards before the start of the WIA program, as they were judged to be too low relative to other states and North Carolina's past

**TABLE 4**  
Departure Points for First Generation JTPA Standards

<i>Program Year</i>	<i>Adult Employment Rate at Termination</i>	<i>Adult Welfare Employment Rate at Termination</i>	<i>Youth Employment Rate at Termination</i>
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

performance (Heinrich 2004). For the most part, states and localities complied with WIA requirements by building yearly increases into the standards.

As the analysis by Heinrich showed, however, this approach failed due to the lack of adjustments for changing economic conditions in the early years of WIA. Two years into the program's operation, 38 states were identified as having failed to achieve at least 80 percent of their performance goals for two consecutive years and were at risk for sanctions. More generally, these findings suggest that the types of formal performance standards adjustments made in the JTPA system to control for factors outside program managers' control are critical to the success of a system intended to promote continuous performance improvements.

## CONCLUSIONS

In this paper, we have drawn from the information economics, contract theory, and public administration literatures to discern simple lessons for the construction of performance standards. We demonstrate the relevance of these lessons in the context of two public programs, the U.S. JTPA and WIA federal job training programs. We find evidence that performance measurement system designers have attempted to "level the playing field" over time to provide equivalent performance incentives across states and localities. Performance standard adjustment methods were established to account for "shocks" that are outside an agent's control and to reduce the risk faced by the agent. Policymakers have also tried to reduce the potential negative distortions due to hidden information.

At the same time, it is not surprising that in a public sector program with multiple principals and political relationships influencing administration, the evidence suggests that these problems were not fully resolved. We identified some negative dynamic properties of the performance measurement system that threaten its sustainability. In both JTPA and WIA, the dynamics of performance benchmarking and the challenges of effectively adjusting performance expectations for external influences beyond program managers' control likely contributed to inefficiencies and generated incentives to influence performance in ways other than increasing effort. Selecting trainees according to observed characteristics associated with their labor market success, limiting the availability of more intensive training services, and demonstrating lower performance early on to allow for performance improvements over time are some examples of strategic behaviors that were unintended by system designers and potentially harmful to the system and program outcomes. In the current system, where the rewards (up to \$3 million in grants) and sanctions (up to a 5% reduction in grants) could have important implications for program functioning, the performance standards should provide appropriate incentives and feedback to operators about the effectiveness of their activities in improving service quality and participant outcomes.

Politicians, along with economists and private sector representatives, have been calling for a more "business-like" administration of government for more than a century, most recently in "reinventing government" and New Public Management

reform initiatives. The use of performance measurement systems and incentives in public sector programs has been a key component of these recent initiatives, although both policymakers and scholars have begun to uncover evidence of their “dark side,” including some of the negative or unintended consequences described in this study (Radin 2000). Our research confirms both the potential of these systems to be effectively managed to promote performance improvements, and the limitations of these systems’ design, which are guided not only by economic theory, but also by political demands and the complexities of representative governance. Although our research doesn’t point to cogent solutions for all of the problems that public sector performance measurement system designers face, we do suggest some specific actions public managers can take to improve these systems, such as the proper incorporation of different types of information into the standard, coordination among multiple principals with conflicting interests, and more careful attention to the dynamic implications of performance measurement. More generally, we also hope that the framework for analysis of these issues that we present might better guide policymakers’ or other scholars’ understanding and consideration of how these systems and public program performance might be improved.

## APPENDIX

Although the DOL allowed states some flexibility in developing standards, most states used the DOL’s adjustment methodology 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  $P_l$ .

$$P_l = \alpha + \beta_1(x_{1l} - \bar{x}_1) + \beta_2(x_{2l} - \bar{x}_2) + \cdots + \beta_M(x_{Ml} - \bar{x}_M) + \varepsilon_l, \quad (1)$$

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  $\varepsilon_l$  is a site-specific error term. Biannually the DOL estimated the coefficients  $\beta$  with the most recent two years of training center-level data using ordinary least squares.  $\beta_m$  expressed the impact of an increase in the factor  $x_{ml}$ , on the outcome  $P_l$ , holding other factors constant. The DOL chose a different set of factors for each performance measure based upon their availability and whether the factors were statistically correlated with the performance outcomes. In addition, political considerations may have played a role.

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

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

<i>A. Local Factors</i>	<i>B. Training Center Factor Values</i>	<i>C. National Averages</i>	<i>D. Difference (B - C)</i>	<i>E. Weights</i>	<i>F. Effect of Local Factors</i>
% 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
Unemployment 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.

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  $\beta$ '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 (i.e., the negative weight implies that on average, compared to men, women produce a lower employment rate at termination).

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.

The 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 880  
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 unemploy-  
ment rate is an example of an influence on the performance outcome that was likely 885  
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 890  
the performance standard in this way may have discouraged cream-skimming.

**NOTE**

1. Both Dixit and Burgess and Ratto (2003) evaluate this literature in the context of incen-  
tive provision inside government organizations.

**REFERENCES**

895

Anderson, K., R. 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.

Q2 Barnow, B. 2000. “Exploring the Relationship between Performance Management and  
Program Impact.” *Journal of Policy Analysis and Management* 19: 118–141. 900

Q2 Barnow, B. and J. 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.

Q2 Bouckaert, G. 1993. “Measurement and Meaningful Management.” *Public Performance and  
Management Review* 17: 31–43. 905

Q2 Brooks, A. 2000. “The Use and Misuse of Adjusted Performance Measures.” *Journal of Policy  
Analysis Management* 19: 323–329.

Q2 Burgess, S. and M. Ratto. 2003. “The Role of Incentives in the Public Sector: Issues and  
Evidence.” *Oxford Review of Economic Policy* 19: 285–300.

Courty, P. and G. Marschke. 1997. “Measuring Government Performance: Lessons from a Fed- 910  
eral Bureaucracy.” *American Economic Review: Papers and Proceedings* 87(2): 383–388.

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

Courty, P. and G. Marschke. 2004. “An Empirical Investigation of Gaming Responses to  
Performance Incentives.” *Journal of Labor Economics* 22: 23–56. 915

Cragg, M. 1997. “Performance Incentives in the Public Sector: Evidence from the Job Train-  
ing Partnership Act.” *Journal of Law, Economics, and Organization* 13: 147–168.

Crewson, P. E. 1997. “Public Service Motivation: Building Evidence of Incidence and Effect.”  
*Journal of Public Administration Research and Theory* 7: 499–519.

Davis, J. H., L. Donaldson and F. D. Schoorman. 1997. “Toward a Stewardship Theory of 920  
Management.” *Academy of Management Review* 22: 20–47.

- Dickinson, K., R. West, D. Kogan, D. Drury, M. Franks, L. Schlichtmann, and M. 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, A. 2002. "Incentives and Organizations in the Public Sector." *Journal of Human Resources* 37: 696–727. 925
- Hatry, H. 1999. "Mini-Symposium on Intergovernmental Comparative Performance Data." *Public Administration Review* 59: 101–104.
- Heckman, J., C. Heinrich, and J. Smith. 2002. "The Performance of Performance Standards." *The Journal of Human Resources* 37: 778–811. 930
- Heckman, J. and J. Smith. 2003. "The Determinants of Participation in a Social Program: Evidence from the Job Training Partnership Act." IZA Discussion Paper no. 798. Bonn, Germany: Institute for Labor Studies. <http://opus.zbw-kiel.de/volltexte/2003/1034/pdf/dp798.pdf>
- Q2 Heckman, J., J. Smith, and N. 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. 935
- Heckman, J., J. Smith and C. Taber. 1996. "What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance in the JTPA Program." Pp. 191–217 in Gary Libecap, ed., *Advances in the Study of Entrepreneurship, Innovation, and Growth*, Vol. 7. Greenwich, CT: JAI Press. 940
- Q2 Heinrich, C. 1999. "Do Government Bureaucrats Make Effective Use of Performance Management Information?" *Journal of Public Administration Research and Theory* 9: 363–393.
- Heinrich, C. 2002. "Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness." *Public Administration Review* 62: 712–725. 945
- Heinrich, C. 2004. "Improving Public-Sector Performance Management: One Step Forward, Two Steps Back?" *Public Finance and Management* 4(3): 317–351.
- Holmstrom, B. 1982. "Moral Hazard in Teams." *Bell Journal of Economics* 13: 324–340.
- Holmstrom, B. and P. Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica* 55(2): 303–328. 950
- Johnston, J. 1987. *The Job Training Partnership Act: A Report by the National Commission for Employment Policy*. Washington, D.C.: U.S. Government Printing Office.
- Q2 Kravchuk, R. and R. Schack. 1996. "Designing Effective Performance-Measurement Systems under the Government Performance and Results Act of 1993." *Public Administration Review* 56: 348–358. 955
- Maltzman, F. and S. Smith. 1994. "Principals, Goals, Dimensionality, and Congressional Committees." *Legislative Studies Quarterly* 19(4): 457–476.
- Q2 Marschke, G. 2003. "Performance Incentives and Organizational Behavior: Evidence from a Federal Bureaucracy." Working paper. University of Albany, State University of New York. 960
- Marshall, M., P. Shekelle, R. Brook, and S. Leatherman. 2000. *Dying to Know: Public Release of Information about Quality of Health Care*. London: Nuffield Trust.
- Miller, G. J. 1992. *Managerial Dilemmas*. Cambridge: Cambridge University Press.
- Perry, J. L. and L. Porter. 1982. "Factors Affecting the Context for Motivation in Public Organizations." *Academy of Management Review* 7: 89–98. 965
- Prendergast, C. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37(1): 7–63.

- Radin, B. 2000. "The Government Performance and Results Act and the Tradition of Federal Management Reform: Square Pegs in Round Holes?" *Journal of Public Administration Research and Theory* 10: 11–35. 970
- Rainey, H. 1982. "Reward Preferences Among Public and Private Managers: In Search of the Service Ethic." *American Review of Public Administration* 50: 374–382.
- Q2 Rubenstein, R., A. Schwartz, and L. Stiefel. 2003. "Better than Raw: A Guide to Measuring Organizational Performance with Adjusted Performance Measures." *Public Administration Review* 63: 607–615. 975
- Q2 Stiefel, L., R. Rubenstein, and A. Schwartz. 1999. "Using Adjusted Performance Measures for Evaluating Resource Use." *Public Budgeting and Finance* 19(3): 67–87.
- Taylor, F. 1911. "Principles and Methods of Scientific Management." *Journal of Accountancy* 12(2): 117–124. 980
- Thompson, J. 1967. *Organizations in Action*. New York: McGraw-Hill Book Company.
- U.S. Department of Labor, Employment and Training Administration. 2001. *2002 Annual Performance Plan for Committee on Appropriations*.
- U.S. Government Accounting Office. 2002. *Improvements Needed in Performance Measures to Provide a More Accurate Picture of WIA's Effectiveness*, GAO Report #02-275. 985
- Q2 Wholey, J. 1999. "Performance-Based Management." *Public Performance and Management Review* 22: 288–307.
- Q2 Wholey, J. and H. Hatry. 1992. "The Case for Performance Monitoring." *Public Administration Review* 52: 604–610.
- Wilson, W. 1887. "The Study of Administration." *Political Science Quarterly* 2(2): 197–222. 990
- Wise, L. 2004. "Bureaucratic Posture: On the Need for a Composite Theory of Bureaucratic Behavior." *Public Administration Review* 64: 669–680.

## ABOUT THE AUTHORS

**Pascal Courty** (pascal.courty@iue.it) is a professor in the Department of Economics and the European University Institute. He received his Ph.D. from the University of Chicago. He joined the EUI in September 2003 from the London Business School where he was associate professor of economics. His research focuses on contract theory with applications to the design of incentives in organisations and to firm pricing policies. 995

**Carolyn Heinrich** (cheinrich@lafollette.wisc.edu) is an associate professor of public affairs at the La Follette School of Public Affairs at the University of Wisconsin–Madison. She is also the Associate Director of Research and Training at the Institute for Research on Poverty. She received her Ph.D. from the University of Chicago. Her research focuses on social welfare policy, public management, and econometric methods to evaluate social programs. 1000 1005

**Gerald Marschke** (marschke@albany.edu) is associate professor in the Department of Economics and Department of Public Administration & Policy at the State University of New York at Albany. He received his Ph.D. from the University of Chicago. His research covers the areas of labor economics, industrial organization, public economics, organizational economics, personnel economics, economics of innovation, and technology policy. 1010