W.E. UPJOHN INSTITUTE
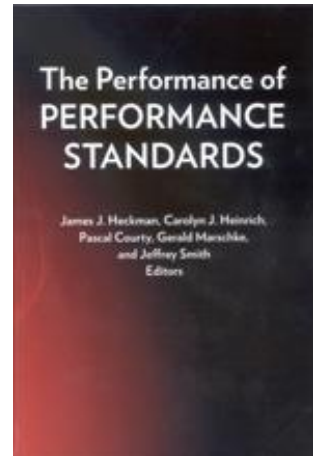FOR EMPLOYMENT RESEARCH

Upjohn Institute Press

Setting the Standards:
Performance Targets and
Benchmarks

Pascal Courty
University of Victoria

Carolyn J. Heinrich
University of Wisconsin

Gerald Marschke
University at Albany, SUNY

The Performance of
PERFORMANCE
STANDARDS

James J. Heckman, Carolyn J. Heinrich,
Pascal Courty, Gerald Marschke,
and Jeffrey Smith
Editors

The Performance of Performance Standards

James J. Heckman
Carolyn J. Heinrich
Pascal Courty
Gerald Marschke
Jeffrey Smith
Editors

2011

The facts presented in this study and the observations and viewpoints expressed are the sole responsibility of the author. They do not necessarily represent positions of the W.E. Upjohn Institute for Employment Research.

5
Setting the Standards

Performance Targets and Benchmarks

Pascal Courty
Carolyn J. Heinrich
Gerald Marschke

A key element in the design of performance measurement and accountability systems is the establishment of appropriate benchmark levels (or standards) of performance to guide the evaluation of program outcomes. Performance benchmarks shape system incentives and influence the responses of public managers and staff operating programs. 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).

In this chapter, we review the literature in information economics, contract theory (see, for example, Dixit [2002] and Prendergast [1999]), and public administration to draw out theoretical implications for the construction of performance standards in public organizations.[1] 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. An important premise of our work is that the method used to construct performance standards can change the way employees behave and influence the internal efficiency of organizations. Focusing in particular on performance benchmarking in U.S. workforce training programs (i.e., JTPA and WIA programs), we assess whether the design of performance standards in these programs is efficient and consistent with basic principles derived from theory.

The exercise of performance assessment clearly serves important functions in public organizations other than promoting efficiency.

Marshall et al. (2000) describes three primary functions: 1) account-ability for public expenditures, 2) the production of comparative information to inform customer choices in public services, and 3) improvement of professional practice and program management. For example, public managers may use performance information to identify best practices and 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 the construction of performance standards and the use of performance data. We discuss these issues to a greater extent in our case study analysis of performance benchmarking in public train-ing programs.

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. There are critics of this model who argue that it overemphasizes the self-seeking behavior of agents and neglects so-cial interactions and motivators. In his classic study of organizations, Thompson (1967), for example, describes the importance of cliques, social controls based on informal norms and status that influence the performance of organizations. Similarly, stewardship theory empha-sizes collective goals and public managers "whose motives are aligned with the objectives of their principals," or who highly value coop-erative behavior even when their interests and those of the principal diverge (Davis, Donaldson, and Schoorman 1997, p. 21). Although we acknowledge the roles of social and cultural norms and the influence of political and personal power relationships as described in these al-ternative 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 per-formance 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 ΔP, 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 multidimensional 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, caseworkers allocate their efforts toward recruiting participants, assessing their training "needs," networking with other social service organizations and 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—are robust to this more realistic assumption.

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.

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 others since the U.S. Government Performance and Results Act (GPRA) of 1993, 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.

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, $P_0 = P(e_0)$.

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 overcompensating 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 was first used in a systematic way (and is still common practice). Taylor (1911) was perhaps its most famous early

proponent. His ideas arose from his experience as a machinist in a steel plant. Miller (1992, p. 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 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, 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 representative 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 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. Counterfactual experiments such as time-and-motion studies, however, are practical only in a few occupations that typically involve manual work. Many public and private sector work situations involve nonmanual work, complex group interactions, and nonstandardized outputs, making experimental studies to construct counterfactual performance 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 stan-

dards and the expected return to the organization from assessing value added more precisely; and of course, in the "real world," non-economic factors—e.g., political goals, legislative requirements, etc.—may also influence standard setting processes. These challenges and trade-offs will become evident in the case study analysis of the JTPA and WIA performance standards that follow. First, however, we review a set of generic problems that any method of 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 different 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 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 standard 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.

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, 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 performance outcome. Consider a more realistic example where the worker produces a number of pieces that depend 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 $P + \Delta$ and the worker's value added is $\Delta P = (P + \varepsilon) - P_0$, where ε 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 that is required to achieve the performance standard will sometimes overperform and other times underperform relative to the standard. In this situation, value added is equal to $\Delta P = \Delta$ when $e = e_0$. Outside shocks do not influence the worker's choice of effort because additional effort still increases expected performance. They do, however, 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 compensation, a risk-averse worker will, and this establishes a first rationale to construct as fair a performance standard as possible. The merit of performance standards will 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 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 hold the worker responsible for effort, *e*, but not for external influences, ε. 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 worker's 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 caseworker responsible for (and reward) efforts made to appropriately assess clients and facilitate better worker-employer matches. But it would be unfair and inefficient to penalize the caseworker (or job-training center) for a lower rate of worker-employer matches if it is due to a declining number of labor market (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 coworkers, compared to a group shock that lowers all workers' performance. The former shock generates different treatments amongst individuals who have essentially behaved identically. As under 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 show 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 that 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 his/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 caseworkers in this program are evaluated in part according to the rate at which their clients secure jobs. Caseworkers observe relevant information about applicants on the likely success of training investments, (e.g., personal motivation and employment barriers). 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 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 employment on their own. As a result the caseworkers may overinvest in easy-to-serve applicants and underinvest 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 information that is publicly known, such as the information about varying productivities of machines. If the caseworker observes the applicants' type, he has an incentive to enroll only easy-to-serve applicants because they produce better 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, 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 he has enrolled, the agent 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 recipient or limited English proficiency as proxies for "hard-to-serve." This would assume that welfare recipients or nonnative 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 nonnative English speakers) are identical; some are easier to serve than others, and the agent observes this information. Again, the agent will be inclined to select a nonrepresentative 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 subpopulations of applicants.

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 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. How-

ever, these agents may not be those who generate the highest returns from the principal's perspective. 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 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 and WIA programs, Congress, the USDOL, and state governments would be the main principals, since these are the actors who directly define the goals and activities of the organization, both through the design of the incentive system, performance standards, and also through other organizational features. But local politicians, private industry council representatives, and other interest groups 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 incentive system, particularly if the interests of the different principals are not aligned, e.g., emphasizing different priorities or outcomes. The agent has to choose how to allocate his/her effort level e 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.

Dixit (2002) proposes an analysis of multiple principals competing noncooperatively for the agent's effort. As expected, the agent will allocate more effort toward the objectives of principals who compensate 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 agent (e_1, e_2, e_3, etc.) toward the achievement of the various outcomes

will be decreasing in 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 he will also insure the agent against failure on dimensions of effort that concern the other principals. If principals choose the level of incentive noncooperatively, the 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 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 calls for 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 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), then it may be optimal to divide up the functions of enrollment and training and to assign each of these activities to two separate agencies.

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 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 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.

Although there is considerable discussion of "multidimensionality" or 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 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 and federal policy directives at the same time, agents might allocate their efforts toward alternative objectives of these principals according to the principals' 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 estimate the production function and 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 his 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 reneging on the contract. Miller (1992, p. 157) likewise recognizes 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 to never change a standard, or more realistically, by committing to strict rules for changing the standard. 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 principal (or 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 earlier, is the introduction of a ratchet effect if higher performance outcomes increase future standards. This method is also unlikely to work well in nonstationary 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 (e.g., akin to the Job Corps Center annual performance rankings). 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 negotiate (agree on) the performance standard. If objective information on the production function is absent and relative performance evaluation is not a viable al-

ternative, this may be the only solution available. This approach requires an environment of mutual trust between the agent and principal(s), i.e., 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, e.g., 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, individual providers of government training have been evaluated by their performance relative to specific, numerical standards. Congress has also legislated important changes in the formulation of these numerical standards, as described in Chapter 4. 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.

Under JTPA, Congress, the USDOL, and state authorities shared in designing and implementing the program's incentive policies. The Labor Department established expected performance levels using a regression-based model with national departure points. States could use the optional department 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 USDOL's parameters (see Chapter 4 and Social Policy Research Associates [1999]). A majority of states adopted these models and used the USDOL-provided performance standards worksheets to determine performance targets (some with modifications).

The WIA program that replaced JTPA in 2000 introduced a new approach to setting performance standards that involves the negotia-

tion of performance targets. States negotiate with the USDOL 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 WIA (USDOL 2001, p. 8).

In our case analysis of the JTPA and WIA performance measurement systems, the USDOL, Congress, and the states constitute multiple principals in the organizational structure, while local implementing authorities (in government entities or training centers) function as the agents, undertaking the business of enrolling, training, and finding employment for the program clients. Table 2.1 in Chapter 2 (p. 22) 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).

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

In general, the USDOL has attempted to address this question through the use of data on past performance. For example, prior to the start of JTPA, the Labor Department collected performance data on outcomes during the final years of the training program that preceded JTPA and used these data to determine the performance standards in the first year of JTPA. Now assume that past performance in a representative training environment gives a distribution of performance, and that differences in performance outcomes are due only to differences in effort. Then if the department 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 nationwide. Thus, the USDOL evaluated a training center's effort, *e*, against an effort level *e'*, that corresponded to the effort level of the training center at the 25th percentile of systemwide 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.

Under WIA, the more systematic approach for setting standards described above was abandoned. With discretion for setting performance standards transferred to the negotiation process between states and localities, the use of past performance information varied widely. The USDOL 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 of these data differed considerably. The various types of data used included the projected national averages for the negotiated standards provided by the Labor Department; federal baseline numbers (available in the federal performance tracking system, i.e., Standardized Program Information Reporting [SPIR] data); unemployment insurance (UI) 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 UI data from earlier periods (PY 1994–1997) combined with USDOL 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.

Is the Playing Field Level?

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 mar-

kets. 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 USDOL 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.

Under JTPA, this method established the 25th percentile only as a starting or departure point. For each training center, the departure point was adjusted using a regression model, taking into account the extent that the training center's characteristics differed from the average training 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.

In the WIA program, the formal performance standards adjustment models were discarded by nearly all of the states (the exceptions being Texas, Maryland, and the District of Columbia). At the same time, the USDOL 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 (e.g., Washington, Nebraska, South Carolina, and others) followed a similar process.

Adjusting for Uncontrollable Risks

In addition to accounting for factors (demographic, economic, or others) known at the time that 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 (ε) 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 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, following the September 2001 World Trade Center terrorist attacks.[2] As unemployment rates were increasing in the first three years of WIA (from 3.94 percent to 5.35 percent, on average) and creating adverse labor market conditions for trainees, the standards for performance achievement in the program were also increasing (from 66.44 percent to 70.94 percent, on average).

Year-to-year variations in job availability typically cannot be anticipated by training centers, much less an economic shock of the magnitude precipitated by the September 11 terrorist attacks. However, 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 GAO report confirmed that WIA program administrators were seriously concerned about their ability to meet performance targets. All state 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 dramati-

Table 5.1 Percent of States Meeting or Exceeding Their Negotiated Performance Standards in PYs 2000–2002

Performance measure/standard	PY 2000	PY 2001	PY 2002
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

SOURCE: Heinrich (2004).

cally, such as the 21 percent decrease in the proportion of states meeting their older youth entered employment rates (see Table 5.1).

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 in the absence of training. This behavior has been called cream skimming (in addition to Chapter 6, see, e.g., Anderson, Burkhauser, and Raymond [1993];

Cragg [1997]; Heckman and Smith [2003]; and Heckman, Smith, and Taber [1996]).

To prevent cream skimming, the USDOL adjusted the JTPA standards for the effects of the characteristics of enrollees on P. As illustrated in Chapter 4, the adjustment method compensated training centers for enrolling persons such as the handicapped who tended to lower posttraining 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 USDOL 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, 2004a) also showed how program managers strategically managed their "trainee inventories" and timed participant program exits to maximize end of the year performance levels. For a more detailed overview of the findings described in this subsection, see Chapters 6 and 7.

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 costs 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 USDOL 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, the GAO (2002) concluded that "the need to meet performance levels may be the driving factor in deciding who receives WIA-funded services at the local level" (p. 14). The GAO report and a subsequent study (Heinrich 2004) describe 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 preprogram earnings were more difficult to replace. A Texas official indicated that even with Texas's 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) estimates 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 is to assess the relationship of local participant characteristics and economic conditions to measured performance and to determine if "adjustments" made in the negotiation process (i.e., 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 char-

acteristics 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 estimates 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 confirm 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 and aftermath of the September 11 terrorist attacks.

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 the 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 with effort, 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, at least if performance improvement or inflation 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 would be used for long enough, performance inflation should eventually stop and about 25 percent of training centers would perform below the standard: an unsatisfactory outcome.

Table 5.2 reports the departure points for a number of the original JTPA performance measures (i.e., the adult employment rate at termination, the adult welfare employment rate at termination, and the youth employment rate at termination). These departure points were consistently set at the 25th percentile of a previous year's distribution of

Table 5.2 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	—	21.4
1985	57.1	—	36.4
1986, 1987	62.4	51.3	43.3
1988, 1989	68.0	56.0	45.0

NOTE: — = data not available.
SOURCE: Courty and Marschke (2004b).

outcomes. As predicted, Table 5.2 shows a general increase in departure points over this period of the early JTPA years. The departure points in 1986–1987 were much higher than those in 1984–1985, 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 USDOL 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 of the states, in fact, attempted to do this. North Carolina, for example, was asked by the USDOL 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 performance (Heinrich 2004). For the most part, though, states and localities complied with WIA requirements by building yearly increases into the standards.

As the analysis by Heinrich shows, 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. The presence of conditions that drive ratchet effects in WIA also suggests that the design of performance incentives may not always follow a strictly economic logic.

CONCLUSIONS

In this chapter, we have drawn from the information economics, contract theory, and public administration literatures to discern basic 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 WIA system, where rewards (up to $3 million in

grants) and sanctions (up to a 5 percent reduction in grants) had potentially important implications for program functioning, the performance standards should have provided 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 businesslike administration of government for more than a century, most recently in the "reinventing government" and New Public Management reform initiatives. The use of performance measurement systems and bonuses 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.

Notes

1. Both Dixit and Burgess and Ratto (2003) evaluate this literature in the context of incentive provision inside government organizations.
2. In New York City alone, it is estimated that about 430,000 job-months and $2.8 billion in wages were lost in the three months following the September 11 attacks (Makinen 2002).

References

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

Burgess, Simon, and Marisa Ratto. 2003. "The Role of Incentives in the Public Sector: Issues and Evidence." *Oxford Review of Economic Policy* 19(2): 285–300.

Courty, Pascal, and Gerald Marschke. 1997. "Measuring Government Performance: Lessons from a Federal Bureaucracy." *American Economic Review* 87(2): 383–388.

———. 2003. "Performance Funding in Federal Agencies: A Case Study of a Federal Job Training Program." *Public Budgeting and Finance* 23(3): 22–48.

———. 2004a. "An Empirical Investigation of Gaming Responses to Performance Incentives" (with P. Courty). *Journal of Labor Economics* 22(1): 23–56.

———. 2004b. "Benchmarking Performance." *Public Finance and Management* 4(3): 288–316.

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

Davis, J. H., L. Donaldson, and F. D. Schoorman. 1997. "Toward a Stewardship Theory of Management." *Academy of Management Review* 22(1): 20–47.

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

General Accounting Office (GAO). 2002. *Improvements Needed in Performance Measures to Provide a More Accurate Picture of WIA's Effectiveness*. Report to Congressional Requesters. GAO-02-275. Washington, DC: GAO.

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." *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. Bonn, Germany: IZA.

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 *Advances in the Study of Entrepreneurship, Innovation, and Economic Growth*, Vol. 7: *Reinventing*

Government and the Problem of Bureaucracy, Gary Libecap, ed. Greenwich, CT: JAI Press, pp. 191–217.

Heinrich, Carolyn J. 1999. "Do Government Bureaucrats Make Effective Use of Performance Management Information?" *Journal of Public Administration Research and Theory* 9(3): 363–393.

———. 2004. "Improving Public-Sector Performance Management: One Step Forward, Two Steps Back?" *Public Finance and Management* 4(3): 317–351.

Holmstrom, Bengt. 1982. "Moral Hazard in Teams." *Bell Journal of Economics* 13(2): 324–340.

Holmstrom, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica* 55(2): 303–328.

Makinen, Gail. 2002. *The Economic Effects of 9/11: A Retrospective Assessment*. Congressional Research Service report, September 27. Washington, DC: Library of Congress. http://www.fas.org/irp/crs/RL31617.pdf (accessed May 28, 2010).

Maltzman, Forrest, and Steven S. Smith. 1994. "Principals, Goals, Dimensionality, and Congressional Committees." *Legislative Studies Quarterly* 19(4): 457–476.

Marshall, Martin, P. Shekelle, R. Brook, and S. Leatherman. 2000. *Dying to Know: Public Release of Information about Quality of Health Care*. London: Nuffield Trust.

Miller, Gary J. 1992. *Managerial Dilemmas*: *The Political Economy of Hierarchy*. New York: Cambridge University Press.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37(1): 7–63.

Radin, Beryl A. 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(1): 11–35.

Social Policy Research Associates. 1999. *Guide to Performance Standards for the Job Training Partnership Act for Program Years 1998 and 1999*. Oakland, CA: SPRA.

Taylor, Frederick Winslow. 1911. "Principles and Methods of Scientific Management." *Journal of Accountancy* 12(2): 117–124.

Thompson, James D. 1967. *Organizations in Action: Social Science Bases of Administrative Theory*. New York: McGraw-Hill Book Company.

U.S. Department of Labor (USDOL). 2001. *2002 Annual Performance Plan for Committee on Appropriations*. Washington, DC: USDOL, Employment and Training Administration.