Robert M.
La Follette School of Public Affairs
at the University of Wisconsin-Madison

Working Paper Series
La Follette School Working Paper No. 2008-022
http://www.lafollette.wisc.edu/publications/workingpapers

The Design and Dynamics of Performance Measurement Systems in the Public Sector

Carolyn J. Heinrich
La Follette School of Public Affairs, University of Wisconsin-Madison

cheinrich@lafollette.wisc.edu

Gerald R. Marschke
Department of Public Administration & Policy and Department of Economics
at the State University of New York at Albany,
National Bureau of Economic Research,
and Institute for the Study of Labor (IZA)

Robert M. La Follette School of Public Affairs
1225 Observatory Drive, Madison, Wisconsin 53706
Phone: 608.262.3581 / Fax: 608.265-3233
info@lafollette.wisc.edu / http://www.lafollette.wisc.edu

The La Follette School takes no stand on policy issues; opinions expressed within these papers reflect the views of individual researchers and authors.

THE DESIGN AND DYNAMICS OF PERFORMANCE MEASUREMENT SYSTEMS IN THE PUBLIC SECTOR

Carolyn J. Heinrich
LaFollette School of Public Affairs
University of Wisconsin-Madison

Gerald Marschke
Department of Public Administration & Policy
and Department of Economics
State University of New York at Albany, NBER and IZA

November, 2008

Preliminary; please do not quote or cite without permission

Abstract

We use the principle-agent model as a focal theoretical frame for synthesizing what we know both theoretically and empirically about the design and dynamics of the implementation of performance measurement systems in the public sector. In this context, we review the growing body of evidence about how performance measurement and incentive systems function in practice and how individuals and organizations respond and adapt to them over time, drawing primarily on examples from performance measurement systems in public social welfare programs (education, employment and training, welfare-to-work and others). We also describe a dynamic framework for performance measurement systems that takes into account strategic behavior of individuals over time, learning about production functions and individual responses, accountability pressures, and the use of information about relationships of measured performance to value-added. Implications are discussed and recommendations derived for improving public sector performance measurement systems.

INTRODUCTION

The use of formal performance measures—based on explicit and objectively defined criteria and metrics—has been a fundamental component of both public and private sector incentive systems since the late 19thcentury. The origins of performance measurement have been traced to the work of industrial psychologists in developing employee rating forms based on psychological traits in the 1800s (Scott, Clothier and Spriegel, 1941), and the US Federal Civil Service has used such individual performance ratings in its performance assessment systems since that time (Murphy and Cleveland, 1995). Moreover, long before the "reinventing government" revolution spearheaded by David Osborne, Ted Gaebler and Al Gore to make government "work better and cost less," scholars such as Woodrow Wilson (1887) were calling for government to become more rational and efficient like the private sector and more accountable to the public.

The typical early performance measurement system debuted in a factory production setting (Taylor, 1911) and was based largely on scientific management principles, that is, promoting the careful analysis of workers' tasks and work arrangements, establishing work procedures according to a technical logic, and setting standards and production controls to maximize efficiency. Through time-and-motion studies, system designers sought to observe an employee's competitive level of effort and to establish a benchmark level (or standard) for performance. With this understanding of the relationship between effort and output, workers could be paid according to a simple model, often linear, that included a base wage per hour plus a bonus rate for performance above the standard.

The basic concepts underlying this simple compensation model—that employees perform better when their compensation is more tightly linked to their effort or outputs, and organizational performance will improve with employee incentives more closely aligned with organizational goals—have been broadly applied in the design of private and public sector performance measurement systems, both historically and currently. The 1910 Taft Commission on Economy and Efficiency, one of the first of a series of major US commissions aimed at improving the performance of government, was significantly influenced by scientific management ideas. Two subsequent commissions—the First (1947-49) and Second (1953-55) Hoover Commissions—were likewise established to "promote economy, efficiency, and improved service in the transaction of the public business...", reflecting the former president's beliefs that "management research technicians" should advise policy and executive agency decisions (The Hoover Commission Report, 1949: xiii; Moe, 1982). Among the series of performance management reforms that followed were the 1960s Planning Programming Budgeting System (PPBS), which featured a "systems analysis" approach to central planning and objective analysis of programs based on research and evaluation, and Zero-Base Budgeting (ZBB), which required managers to narrowly define and measure progress toward financial, technical, and strategic performance goals. In fact, ZBB has recently been employed by two large states, Florida and Texas, that are widely recognized as leaders in performance management, although Moynihan (2008) criticizes their adoption of ZBB as resulting in ineffective use of performance information and bizarre consequences.

Bertelli and Lynn (2006: 28-9) suggest that the early technical focus of public sector performance measurement systems was congruent with a prevailing "scientism in political science," that is, with an orientation toward descriptive analysis of formal governance structures and processes rather than attention to the dynamics of system incentives and their consequences. Indeed, public managers still had little leeway or authority following these reforms to base employees' pay on their performance. Rigid systems of job and position classification and pay raises based on time in service rather than performance contributed to significant disparities between public and private sector managers' ability to correct or dismiss employees based on their performance (Rainey, 1983). The U.S. Civil Service Reform Act of 1978, widely regarded as the most far-reaching administrative reform legislation since 1883, sought to remedy this by allowing performance-contingent pay and increasing managers' latitude for rewarding performance. The ensuing development and testing of "pay for performance" systems for both personnel management and public accountability is ongoing, although research suggests that public sector applications have to date met with limited success, primarily due to inadequate performance evaluation methods and underfunding of data management systems and rewards for performance (Rainey, 2006; Heinrich, 2007).

In reality, with each successive reform, we have come to more fully appreciate that the conditions and assumptions under which the simple, rational model for a performance measurement and incentive system model works—organizational goals and production tasks are known, employee efforts and performance are verifiable, performance information is effectively communicated, and there are a relatively small number of variables for managers to control — are stringent, and in the public sector, rarely observed in practice. Rather, many private and public sector settings are more likely to exhibit "open" system attributes, i.e., containing "more variables than we can comprehend at one time" and some variables "subject to influences we cannot control or predict," along with a wider range of work motivations and organizational cultures (Thompson, 1967:6; Moynihan, 2008).

The fact that many private and public sector production technologies typically involve complex, non-manual work, multiple principals and group interactions, political and environmental influences and interdependencies, and non-standardized outputs makes the accurate measurement of performance and construction of performance benchmarks (through experimental studies, observational methods, or other approaches) both challenging and costly.

As Dixit (2002) observed, these richer "real world" circumstances often make it inappropriate to rely heavily on incentives linked in simple ways to output measures that inexactly approximate government performance. Yet, because of their ease of use and the significant costs associated with developing and implementing more sophisticated measures and more intricate contracts or systems of incentives, simple incentives built around inexact and incomplete measures are still widely used for inducing work and responsible management. Indeed, Bertelli and Lynn (2006:70) assert that scientific management—in the form of performance measurement, cost-benefit analysis and other information technologies—is "as much in vogue as it ever was" in public administration.

Recent public sector reforms aimed at improving government performance, such as the U.S. Government Performance and Results Act (GPRA) of 1993 and the Program Assessment Rating Tool (PART) used by the Office of Management and Budget, likewise reflect this simplistic approach to motivating and measuring performance. Regardless of the goals or intricacy of their work, every federal agency is required to establish performance goals and measures and to provide evidence of their performance relative to targets in annual reports to the public. Furthermore, distinct from earlier public sector performance measurement systems are the focus on measuring *outcomes* and the greater demands for external visibility and accountability of these performance measurement activities, along with higher stakes attached to performance achievements through increasing use of performance-contingent funding or compensation, organization-wide performance bonuses or sanctions and competitive performance-based contracting (Heinrich, 2007). In his comprehensive study of performance management reforms, Moynihan (2008: 4-5) asks: "How important are these performance management reforms to the actual management of government?" He answers: "It is only a slight exaggeration to say that we are betting the future of governance on the use of performance information."

As a result of these ongoing and growing efforts to design and implement performance measurement systems in complex organizational settings, we have been able to compile a growing body of information and empirical evidence about how these incentive systems function in practice and how individuals and organizations respond and adapt to them over time. The primary goals of this Policy Retrospectives piece are to synthesize what we know both theoretically and empirically about the design and dynamics of the implementation of public sector performance measurement systems and to distill important lessons that should inform performance measurement and incentive system design going forward.

Recognizing that the research literature on performance measurement is diverse in both disciplinary (theoretical) and methodological approach, it would be impractical to try to present an all-inclusive review in a single article such as this. Rather, we choose a focal theoretical frame—the principle-agent model—that we argue is the most widely applied across disciplines in studies of performance measurement systems, and we draw in other theories to elaborate on key issues. In addition, our review of the literature focuses primarily on studies that produce empirical evidence on the effects of performance measurement systems, and in particular, on systems that objectively measure performance outcomes. We discuss examples of performance measurement systems, largely in public social welfare programs (education, employment and training, welfare-to-work and others), to highlight important challenges in the design and implementation of performance measurement systems, both in the context of a broader theoretical framework and in their application. We also acknowledge that there is a very rich vein of qualitative studies that incisively investigate many of these issues (see, for example, Radin, 2006; Frederickson and Frederickson, 2006; Moynihan, 2008), but we regret that it is beyond a manageable scope of this review to fully assess their contributions to our knowledge of performance measurement system design and implementation.

In the section that follows, we briefly describe alternative conceptualizations of performance measurement systems and then focus the discussion on key elements of a commonly used static, principal-agent framework for informing the development of performance incentive systems. We also highlight the limitations of particular features of these models for guiding policy design and for our understanding of observed individual and organizational responses to performance measurement requirements. We next consider the implications of a dynamic framework for performance measurement system design that takes into account the strategic behavior of agents over time, learning about production functions and agent responses, and the use of new and/or better information about the relationship of measured performance to value-added. We are also guided throughout this review by a basic policy question: do performance measurement and incentive systems work as intended in the public sector to more effectively motivate agencies and employees and improve government outcomes? In the concluding section, we discuss recommendations for improving public sector performance measurement systems and for future research in this area.

THEORETICAL MODELS FOR PERFORMANCE MEASUREMENT AND INCENTIVE SYSTEMS

Theoretical models that have been applied in the study of performance measurement and incentives range from a very basic “logical model” of “relevant factors” in a performance measurement system (Hatry, 1999) to more elaborate frameworks that attempt to account for multiple principals and levels of influence, complex work technologies and group interactions, political and environmental influences and interdependencies, and other factors. For example, Hatry’s simple model links inputs, activities, outputs and outcomes (intermediate and end) and describes relationships among these different types of performance information. Although his model does not formally identify the influence of context or relationships among performance measures at different levels, he does call for public managers to gather explanatory information’ along with their performance data—from qualitative assessments by program personnel to in-depth program evaluations that produce statistically reliable information—to interpret the data and identify problems and possible management or organizational responses.

In an approach that more clearly diverges from the scientific logic, Moynihan (2008: 95) articulates an “interactive dialogue model of performance information use” that describes the assembly and use of performance information as a form of “social interaction.” The basic assumptions of his model—that performance information is ambiguous, subjective and incomplete, and context, institutional affiliation, and individual beliefs will affect its use—challenge the rational suppositions underlying most performance management reforms to date. While performance management reforms have emphasized what he describes as the potential “instrumental benefits” of performance management, such as increased accountability and transparency of government outcomes, Moynihan suggests that elected officials and public managers have been more likely to realize the “symbolic benefits” of creating an impression that “government is being run in a rational, efficient and results-oriented manner” (p. 68). Moynihan argues convincingly for the cultivation and use of strategic planning, learning forums, dialogue routines and related approaches to encourage an “interactive dialogue” that supports more effective use of performance information and organizational learning. He also asserts that performance management system designers and users need to change how they perceive successful use of performance information, namely, that performance information systems and documentation of performance are not an “end” but rather a means for engaging in policy and management change.

Another heuristic model for government performance analysis, set forth by Lynn, Heinrich and Hill (2001), is intended to aid researchers in explicitly modeling complex, multilevel governance relationships and individual and organizational responses to performance incentives, as well as their implications for government outcomes. Their model delineates a hierarchy of relationships among: i) legislative and political choice (i.e., responsibilities for implementing public law), ii) governance structures, iii) management strategies, iv) core technologies and organizational functions, v) program outcomes, and vi) client/citizen assessments.[1] This framework has been applied in a number of studies of government performance, including public school performance, welfare and job-training program outcomes, health care services outcomes, and the contracting out of government services, often in conjunction with multilevel statistical modeling strategies that aim to identify causal relationships within and across hierarchical levels of government (Brudney et al., 2005; Ewalt and Jennings, 2004; Heinrich and Fournier, 2005; May and Winter, forthcoming). Bloom, Hill, and Riccio (2005), for example, conducted a multilevel re-analysis of data from a multi-site evaluation of the Job Opportunities and Basic Skills (JOBS) program, an early welfare-to-work program. In analyzing the effects of management strategies, core services, client characteristics and external factors such as economic environment on performance (i.e., the earnings of welfare-to-work clients) across local offices, they found that management choices and practices related to goals, client-staff interfaces, and service strategies had substantive, statistically significant effects on client outcomes and local office performance.

The principal-agent model that we focus on this paper lies somewhere between a simple

[1] It is also the explicit intent of this framework to encourage researchers to recognize and acknowledge any influential factors at levels of government that they (perhaps necessarily) ignore in their models and the likely implications of these unmeasured factors for accurately assessing performance.

logical model and a more elaborate governance model in its description and framing of relationships among factors that influence performance. We argue that not only is it the most widely applied in the study of performance measurement and incentive systems across disciplines, but it also yields important insights that are fundamental to the design and implementation of any performance measurement and incentive system. For example, it provides a cogent framework for addressing one of the most common problems in assessing agent effort and performance and aligning the incentives of multiple system actors (organizations or individuals), that is, asymmetric information. The principal-agent model is also particularly informative for analyzing how organizational structures, performance measurement system features, and management strategies influence the responsiveness of agents to the resulting incentives. In this regard, it appropriately fits within the broader governance framework described above.

A Static Framework for Performance Measurement and Incentive System Design

We begin with the multitasking model (Holmstrom and Milgrom, 1991; Baker, 1992), typically presented with a risk neutral employer and a risk averse worker,[2] to outline a static framework for the design of incentive systems in public and private organizations. An organization (the principal) hires a worker (the agent) to perform a set of tasks or actions. The organization cannot observe the worker's actions, nor can the organization precisely infer it from

[2] The principal's risk preferences are assumed different from the agent's because the model is frequently used in contexts where the principal is the stockholder of a for-profit firm and the agent is a worker in the firm. The firm's stockholders may have a comparative advantage in risk management because they can more easily spread their wealth among different assets, while workers may have most of their wealth tied up in their (non-diversifiable) human capital, limiting their ability to manage their risk. Moreover, persons selecting into occupations by their ability to tolerate risk implies that the owners of firms will be less risk-averse than firms' workers. In the public sector, the goal of the organization is not profit but policy or program value-added, and the principal is not the owner of a firm but a public sector manager or a politician. The public sector principal may not as easily diversify away the risk of managerial or policy failures by purchasing assets whose risk will offset policy risk (a point made by Dixit, 2002, p. 5). Thus the assumption that the principal and agent have different risk attitudes in public sector applications of the model may not be as justifiable.

measures of organizational performance. Following the formulation of the multitasking model in Baker (2002), let V be the total value of the organization. If the organization is a firm, V is the present value of the net income stream to the firm's owners. Alternatively, if the organization is in the public sector, V is the value generated by the organization for the citizens it serves net of the tax contributions to its production. The relationship between V and the government employee's actions is

$$V(\mathbf{a}, \varepsilon) = \mathbf{f} \cdot \mathbf{a} + \varepsilon \qquad (1)$$

where **a** is a vector of actions that the worker can take to affect V, **f** is the vector of associated marginal products, and ε is an error term (mean zero, variance σ_{ε}^{2}) that capture the effects of uncontrollable events or environmental factors that impact value (including some political variables). The literature typically assumes the organization does not observe **a** and ε, but does know the functional relationship among the variables, including **f**, and the distribution from which ε is drawn.

The assumptions that the principal cannot detect and therefore reward effort, and that effort is costly to the agent, ensure that the agent will shirk if paid a flat salary. Because it is directly affected by **a**, if V were observable, the organization could base the employee's compensation on V, thereby creating an incentive to exert effort. In the public sector, the value-added of programs is occasionally assessed through the use of randomized experiments, but more often than not, V is not observed.[3] However, there are often available one or more performance measures that reveal information about the employee's actions. Assume the existence of a performance measure P such that

$$P(\mathbf{a}, \varphi) = \mathbf{g} \cdot \mathbf{a} + \varphi \qquad (2)$$

[3] To be clear, we are using the term "value" or "value-added" in an impact evaluation sense, where the value added of a program or of an employee to a given organizational effort is assessed relative to the outcome in the absence of the program or the employees' efforts, i.e., the counterfactual state that we do not observe.

where φ is an error term (mean zero, variance σ_{φ}^{2}) that captures the effects of uncontrollable events or environmental factors on the performance outcome. As with V, it is assumed that the organization doesn't observe φ but does know the functional relationship among the variables, including **g,** and the distribution from which φ is drawn. Note that while they may share common components and may be correlated, φ and ε are generally different.

Relationship between performance and value-added and choice of performance measures

The observation that **g** is not necessarily equal to **f**—that is, the effect of an action on performance (P) may be different than its effect on value (V)—and that this difference should be considered in constructing a wage or work contract, is a key insight of the research of Baker (1992) and Holmstrom and Milgrom (1991). In other words, **f** and **g** may differ if some actions in **a** that influence P do not affect V, or worse yet, if some actions that increase P simultaneously reduce V. An example of the former is a teacher providing her students with (or filling in) the answers to a state-mandated test (Jacob and Levitt, 2003). Differences in **f** and **g** may also induce the worker to exert the right kinds of effort but in the wrong quantities—e.g., a teacher "teaching to the test."

By linking the worker's compensation to P, the organization may increase the worker's attention to V-increasing actions. One of the simplest compensation schemes is a linear contract: the agent receives a salary independent of the agent's effort and performance, plus a sum that varies with performance. The variable part is a piece rate—a fixed amount of compensation per unit of performance—times P. The higher the piece rate, the greater is the fraction of the agent's compensation that stems from performance and the stronger the incentive to exert effort. On the other hand, the higher the piece rate, the greater the influence of factors outside the agent's control on compensation, and thus, the greater the risk taken on by the agent.

Baker's (1992) model shows how the organization should navigate this trade-off, which has been studied extensively in the literature (Holmstrom, 1982). The sensitivity of an agent's compensation to performance should be higher the more beneficial (and feasible) is an increase in effort. Increasing the intensity of performance incentives is costly to the organization, because it increases the risk the agent bears, and correspondingly, the compensation necessary to retain her. Thus, raising the incentive intensity makes sense only if the added effort is consequential to the value of the organization. Everything else equal, the sensitivity of an agent's compensation to performance should be lower the higher the agent's risk aversion and also the noisier the outcome σ_{φ}^{2}. That is, the less informative P is about the agent's effort, the less the principle should rely on it as a signal of effort.

Incentives should also be more intense the more responsive is the agent's effort to an increase in the intensity of incentives. That is, incentives should be used in organizations for agents who are able to respond to them. In some government bureaucracies, workers are highly constrained by procedural rules and regulations that allow little discretion. Alternatively, in organizational environments where agents have wide discretion over how they perform their work, agents may respond to performance incentives with innovative ways of generating value for the organization. Imposing performance incentives on agents who have little discretion or control needlessly subjects them to increased risk of lower compensation.

In addition, Baker's model introduces another trade-off that organizations face, namely, the need to balance the effort elicited by a high piece rate (based on P) with the possible distortions it may cause. The sensitivity of an agent's compensation to performance should be smaller the more misaligned or distortionary the performance measure with respect to the value of the organization, that is, the less aligned are **f** and **g**. A key implication of this model is that the optimal incentive weight does not depend on the sign and magnitude of the simple correlation between V and P, which is influenced by the relationship between external factors that affect the performance measure and those that affect the organization’s value. Rather, what matters is the correlation between the effects of agent actions on outcomes and on the organization’s value.

For example, from the perspective of the policymaker or taxpayer who would like to maximize the value from government dollars spent on public programs, we want to choose performance measure(s) so that the effects of agent actions on measured performance are aligned with the effects of those same actions on value. That is, for a given performance measure, we would like to understand how closely **g** and **f** align, but since we frequently don’t know **g** and may likewise have little information about **f**, this is difficult to realize in practice. Empirical research on this issue has focused primarily on statistically estimating measures of association between P and V (using data from randomized experiments), which tells us little about the attributes of P that might make it a useful measure, such as the degree of noise and its alignment with V.

The choice of performance measures is further complicated by the fact that organizations typically value performance across a set of tasks, and that the workers’ efforts are substitutes across tasks. The optimal contract will then place weight on the performance measure corresponding to each task, and the agent will respond by exerting effort on each task (Holmstrom and Milgrom, 1991). As is often the case, however, suppose that for only a subset of the tasks does there exist a performance measure that is informative. Holmstrom and Milgrom show that if an organization wants a worker to devote effort to each of several tasks or goals, then each task or goal should earn the worker the same return to effort at the margin. Otherwise the worker will devote effort only to the task or goal that has the highest return to effort. The implication is that in the presence of productive activities for which effort cannot be measured (even imprecisely), weights on measurable performance should be set to zero. Or as Holmstrom and Milgrom (1991: 26) explained, “the desirability of providing incentives for any one activity decreases with the difficulty of measuring performance in any other activities that make competing demands on the agent’s time and attention,” which is perhaps why at least in the past, the use of outcome-based performance measures in public social welfare programs was relatively rare.

It has been suggested that in such organizational contexts, subjective performance evaluation by a worker's supervisor allows for a more nuanced and balanced appraisal of the worker's effort. Furthermore, subjective evaluation may make evaluation based on objective performance measures more effective (Baker, Gibbons, and Murphy, 1994). Subjective evaluation is itself imperfect, however, as it produces measures that outside parties cannot verify, and subjective measures are also subject to distortion. At the same time, it has also been argued by Stone (1997: 177) and others (Moynihan, 2008) that *all* performance information is subjective, as any number can have multiple interpretations depending on the political context, that is, given that they are “measures of human activities, made by human beings, and intended to influence human behavior.”

Employee motivations and choice of action

One might ask why employees would take actions that increase only measured performance (P), with little effect on value (V). Using public employment and training programs as an example, consider the technology of producing V, such as a real increase in individuals’ skills levels, compared to what is required to arrange a job entry for a program participant (one of the long-used performance measures in these programs). Based on our prior research in this area (Courty and Marschke, 2004; Heckman, Heinrich and Smith, 2002), we suggest that it will require greater investments of resources and effort on the part of employees, such as more intensive client case management and skills training provided by more experienced program staff, to affect V compared to the level of effort required to increase P in this case. Given that V does not enter into the employee's compensation contract, the employee will have little incentive to choose actions in **a** that incur a higher cost to him (to produce V), particularly if these actions do not correlate strongly with actions that affect P.

At the same time, there is a growing interdisciplinary literature that suggests that individuals in organizations may be motivated to exert effort in their work by something other than the monetary compensation they receive. In the public administration literature, empirical tests of the theory of "public service motivation" provide support for the tenet that some bureaucrats are motivated by "an ethic to serve the public," i.e., acting because of their commitment to the common good rather than self-interest (Perry & Porter, 1982; Rainey, 1982; Perry & Wise, 1990; Houston, 2006). Prendergast (2007) refers to these employees who care about their work (and less so about the monetary rewards) as "intrinsically" motivated. And stewardship theory, which derives from psychology, argues that some employees are not motivated by individual goals, but rather align their interests and objectives with those of the principal (Davis, Schoorman, & Donaldson, 1997; Van Slyke, 2007). In other words, organizational stewards perceive that their personal needs and interests are met by achieving the goals of the organization.

The implications of intrinsic, stewardship or public-service motivations among employees are that they derive utility (i.e., intrinsic rewards) from work and thus will exert more effort for the same (or a lower) level of monetary compensation, and through their identification with the goals of the principal or organization, they are also more likely to take actions that are favorable to the interests of the principal (Akerlof and Kranton, 2005; Murdock, 2002; Prendergast, 2007; Van Slyke, 2007; Francois and Vlassopoulos, 2008). In other words, if value-added (V) is the goal of the organization (e.g., increasing clients' skills levels in an employment and training agency) and is communicated as such by the principal, intrinsically motivated employees should be more likely to work harder (with less monitoring) toward this goal and should be less likely to take actions that only increase P and more likely to take actions that only increase V. The returns to exerting effort in activities in the correct quantity to maximize V are greater for the intrinsically motivated agent, and since he is less interested in monetary pay, he will be less concerned about increasing P to maximize his bonus.

The findings of Heckman et al. (1996, 1997) and Heinrich (1995) in their studies of public employment and training centers under the U.S. Job Training Partnership Act (JTPA) program aptly illustrate these principles. They showed that the hiring of case workers in Corpus Christi, Texas who exhibited strong preferences for serving the disadvantaged, in line with a stated goal of JTPA to help those most in need, likely lowered service costs (i.e., wage costs) in Corpus Christi. Alternatively, an emphasis on client labor market outcomes in a Chicago area agency that was exceptional in its attention to and concern for meeting the federal performance standards appeared to temper these public service motivations. With performance expectations strongly reinforced by administrators and in performance-based contracts, case workers' client intake and service assignment decisions were more likely to be made with attention to their effects on meeting performance targets and less so with concern for who would benefit most from the services (the other basic goal of the JTPA legislation).

In the undesirable case in which both principal and agent motivations and the effects of agent actions on P and V do not align, the consequences for government performance can be disastrous, as Dias and Maynard-Moody (2007) showed. They studied workers in a for-profit subsidiary of a national marketing research firm that shifted into the business of providing welfare services. As they explained, requirements for meeting contract performance (job placement) goals and profit quotas generated considerable tensions between managers and workers with different philosophies about the importance of meeting performance goals vs. meeting client needs. They noted that the easiest way to meet contract goals and make a profit was to minimize the time and effort line staff devoted to each client, as individualized case management requires "detailed, emotional interaction with clients, taking time to discover their areas of interest and identifying the obstacles preventing them from accomplishing their individual goals," which distracts from the goal of immediate job placement. In the face of these pressures, managers had little incentive to go beyond minimal services, and they consequently assigned larger caseloads to front-line staff to allow less individual time with clients. As Dias and Maynard-Moody reported, however, the differences in views of the intrinsically motivated staff (focused on V) and managers concerned about P triggered an "ideological war" that subsequently led to "enduring and debilitating organizational problems" and failure on both sides to achieve the program goals.

In addition to monetary and intrinsic incentives, empirical research also suggests that agents respond strongly to non-pecuniary incentives, which in the public sector, most frequently involve the potential for public recognition for performance (Bevan and Hood, 2006; Walker and Boyne, 2006; Heinrich, 2007). Incentive power is typically defined in monetary terms as the ratio of performance-contingent pay to fixed pay, with higher ratios of performance-contingent, monetary pay viewed as "higher-powered," that is, offering stronger incentives (Asch, 2005). Pecuniary-based incentives are scarcer in the public sector, however, in part because of insufficient funds for performance bonuses but also reflecting public aversion to paying public employees cash bonuses. Although recognition-based incentives might have quite different motivational mechanisms than monetary compensation, research suggests they may produce comparable, or in some cases, superior results. For example, recognition may bolster an agent's intrinsic motivation or engender less negative reactions among co-workers than pecuniary awards (Frey and Benz, 2005). In Heinrich's study of performance bonus awards in the Workforce Investment Act, she observed that performance bonuses paid to states tended to be low-powered, both in terms of the proportion of performance-contingent funds to total program funding and the lack of relationship between WIA program employees' performance and how bonus funds were expended by the states. Yet WIA program employees appeared to be highly motivated by the non-monetary recognition and symbolic value associated with the publication of WIA high performance bonus system results. In other words, despite the low or negligible monetary bonus potential, WIA employees still cared greatly about their measured performance outcomes and actively responded to WIA performance system incentives.

Job design and performance measurement

Holmstrom and Milgrom (1991) explore job design as a means for controlling system incentives in their multitasking model in which the principal can divide responsibility for work tasks among agents and determine how performance will be compensated for each task. Particularly in the presence of agents with differing levels or types of motivation, the principal could use this information to group and assign different tasks to employees according to the complexity of tasks and how easily outcomes are observed and measured. Employees who are intrinsically motivated to supply high levels of effort and are more likely to act as stewards (identifying with the goals of the principal) could be assigned tasks that require a greater level of discretion and/or are more difficult to monitor and measure.[4] For tasks in which performance is easily measured, the incentives provided for working hard and achieving high levels of performance should be stronger.

We consider the potential of job design to address an ongoing challenge in the performance measurement systems of public programs—the selective enrollment of clients

[4] Job assignment by preferences of course assumes that the principal can observe the agent's preference type.

and/or limitations on their access to services that is intended to increase measured performance, P, even in the absence of any value added by the program. In the U.S. JTPA program discussed above, job placement rates and wage at placement performance were straightforward to measure but also easy to "game" by enrolling individuals with strong work histories (e.g., someone recently laid off from a well-paying job). Partly in response to this problem, the JTPA program subsequently added separate performance measures for "hard-to-serve" groups, and its successor, the Workforce Investment Act (WIA) program, added an earnings change measure. As we discuss further below, however, the addition of these measures did not discourage strategic enrollments, but rather, they primarily changed the groups who were affected by this behavior.

Since it is difficult to dissuade this type of behavioral response, another possibility would be to separate work tasks, such as enrollment, training and/or job placement activities performed by groups of employees. Prendergast's (2007) research suggests that in this case, workers who are more intrinsically motivated should be assigned the tasks of training (i.e., increasing clients' skills levels), as training effort is more difficult to evaluate and is more costly to exert. No incentives or low-powered incentives should be tied to the training activity outputs. Alternatively, extrinsically motivated employees could undertake activities for which performance is easier to measure, such as job development activities and enrollment tasks that specify target goals for particular population subgroups; with their performance more readily observed, higher-powered incentives could be used to motivate these employees. And ideally, the measured performance of employees engaged in job placement activities would be adjusted for the skills levels achieved by clients in the training component of the program. Under WIA, the more intensive training services are currently delivered through individual training accounts (ITAs) or vouchers by registered providers, which in principle, would facilitate a separation of tasks that could make a differential allocation of incentive power feasible. However, to realize the incentive system design described above, the registration of providers would have to be limited to those with primarily intrinsically motivated employees, which would effectively undermine the logic of introducing market-like incentives (through the use of vouchers) to encourage and reveal agent effort.

In the area of welfare services delivery, there has recently been more systematic study of the potential for separation of tasks to influence employee behavior and program outcomes. In her study of casework design in welfare-to-work programs, Hill (2006) explored empirically the implications for performance of the separation of harder-to-measure, core casework tasks—i.e., client needs assessments, the development of employability plans, arranging and coordinating services, and monitoring of clients' progress—from other caseworker tasks that are more readily measured, such as the number of contacts with employers for job development or the number of jobs arranged. Her central hypothesis was that separating the stand-alone casework tasks (from the core tasks) into different jobs would lead to greater program effectiveness, as measured in terms of increasing client earnings and reducing AFDC receipt. Hill found support for her hypothesis "consistent with the incentive explanation from multitask principal agent theory—that is, staff are better directed toward organizational goals when unmeasurable and measurable tasks are grouped in separate jobs" (p. 278). The separation of casework tasks contributed positively to welfare-to-work program client earnings, although she did not find statistically significant associations with reductions in client receipt of AFDC benefits.

More often than not, though, it is simply not pragmatic to separate work tasks in complex production settings, and some administrative structures in the public sector may only allow for joint or "team production." In fact, there has been a shift toward the use of competitive monetary bonuses to motivate and reward performance at higher (aggregate) levels of government organization, such as job training centers under JTPA and the Job Corps program, schools or school districts in public education systems, and states in the WIA and Temporary Assistance for Needy Families (TANF) programs. In these and other public programs, aggregate performance is calculated from client-level information, and performance bonuses are awarded based on rankings or other formulae for assessing the achievement of performance goals by the agency, state or other aggregate unit. As Holmstrom (1982) explicated, the use of relative performance evaluation such as this will be valuable if one agent or group's outcomes provide information about the state uncertainty of another, and only if agents face some common uncertainties.

For example, the postmaster general's office supervises thousands of local post offices, each providing the same kinds of services but to different populations of clients. The performance of a local office's peers will contain information about its effort, given that the performance of each is a function of not only the effort exerted, but also of external, random factors. Some of these external factors will be idiosyncratic, such the weather, while others are shared, such as federal restrictions on postal rate charges. Relative performance evaluation—basing an agent's performance on how it compares to his peers—works by differencing out the common external factors that affect all agents' performance equally, which reduces some of the risk the agent faces and allows the principal to raise the incentive intensity, and subsequently, agents' effort levels.[5] It is one example of Holmstrom's (1979) "informativeness principle," that is: the effectiveness of an agent's performance measure can be improved by adjusting the agent's performance by any secondary factor that contains information about the measurement error. Another example is the use of past performance information to learn about the size of the random

[5] While relative performance evaluation among agents within an organization may lead to a more precise estimate of agent effort, it may also promote counterproductive behavior in two ways. First, pitting agents against each other can reduce the incentive for agents to cooperate. In organizations where the gains from cooperation are great, the organization may not wish to use relative performance evaluation, even though it would produce a more precise estimate of agent effort. Second, pitting agents against one another increases the incentive for an agent to sabotage another's performance. These behaviors would militate against the use of tournaments in some organizations.

component and to reduce the noise in the measure of current effort, an approach that is currently being used in many school districts to measure the "value-added" of an additional year of schooling on student achievement (Harris and Sass, 2006; Koedel and Betts, 2008; Rothstein, J., 2008).[6]

Both of these examples of the application of the "informativeness principle" were reflected in the U.S. Department of Labor's efforts under JTPA to establish expected performance levels using regression-based models. States were allowed to modify these models in calculating the performance of local service delivery areas, taking into account past performance and making adjustments for local area demographic characteristics and labor market conditions that might be correlated with performance outcomes. For example, performance expectations were lower for training centers located in relatively depressed labor market areas compared to those located in tighter labor markets. The aggregate unit for performance measurement was changed under WIA, however, to the state level; the states' aggregate performance relative to standards set for seventeen different outcomes is calculated, and bonuses are awarded based on states' performance levels, their relative performance (compared to other states), and performance improvements. In accord with Holmstrom's observations, the shift to a state-level performance measurement system has been problematic (see Heinrich, 2007). The local-level variation that was useful for understanding the role of uncertainties or uncontrollable events in performance outcomes is lost in the aggregation, and information gathered about performance is too far from the point of service delivery to be useful to program managers in understanding how their performance can be improved.

[6] Using an assessment period to create the performance standard creates an incentive for the agent to withhold effort. The agent will withhold effort because he realizes that exceptional performance in the assessment period will raise the bar and lower his compensation in subsequent periods, everything else equal. This is the so-called *ratchet* problem. A way to make use of the information contained in past performance while avoiding the ratchet problem is to rotate workers through tasks. By setting the standard for an agent based on the performance of the agent that occupied the position in the previous period, the perverse incentives that lead to the ratchet problem are eliminated.

Dynamics Issues in Performance Measurement and Incentive Systems

As discussed above, the multi-tasking literature commonly employs a static perspective in modeling performance measurement design, applicable to both the public and private sectors.[7] In practice, however, "real-world" incentive designers and policy makers typically begin with an imperfect understanding of agents' means for influencing measured performance and then learn over time about agents' behavior and modify the incentives agents face accordingly. For example, achievement test scores have long been used in public education by teachers, parents and students to evaluate individual student achievement levels and grade progression, based on the supposition that test scores are accurate measures of student mastery of the subject tested, and that increases in test scores correlate with learning gains. More recently, however, under federal legislation intended to strengthen accountability for performance in education (No Child Left Behind Act of 2001), public schools are required to comprehensively test and report student performance, with higher stakes (primarily sanctions) attached to average student (and subgroup) performance levels.

Chicago Public Schools was one of the first school districts to appreciably increase the stakes for student performance on academic achievement tests, beginning nearly a decade before the pressures of NCLB came to bear on all public schools. The early evidence from Chicago on the effects of these new systems of test-based accountability suggests some responses on the part of educators that were clearly unintended and jeopardized the goal of producing useful information for improving educational outcomes. In his study of Chicago Public Schools, Brian Jacob (2005) detected sharp gains in students' reading and math test scores that were not predicted by prior trajectories of achievement. His analysis of particular test question items showed that test score gains did not reflect broader skills improvement but rather increases in

[7] For an exception, see Courty and Marschke (2003, 2007b).

test-specific skills (related to curriculum alignment and test preparation), primarily in math. In addition, he found some evidence of shifting resources across subjects and increases in student special education and retention rates, particularly in low-achieving schools. And with Steven Levitt (2003), he produced empirical evidence of outright cheating on the part of teachers, who systematically altered students' test forms to increase classroom test performance. The findings of these and related studies call into question the use of student achievement test scores as a primary measure to guide efforts to improve the performance of public schools, although critics of current test-based accountability systems have yet to identify any alternative (and readily available) measures that are less noisy or prone to distortion.

Other examples suggest that incentive designers who discover dysfunctional behavior will often choose to replace or modify performance measures rather than eliminate incentive systems altogether. In the JTPA and WIA programs, policymakers added new performance measures (and discarded others) over time to reduce case workers' focus on immediate outcomes (job placement and wages at the time of program exit). There are now seventeen performance measures in use in the WIA program, including measures of employment retention, skill attainment and "credential" rates, and earnings changes. JTPA's original set of performance measures also included cost measures—based on how much employees spent to produce a job placement—to encourage program managers to give weight to efficiency concerns in their decision-making. However, eight years after their introduction, JTPA officials phased them out in response to research and experience showing that the cost standards were limiting the provision of longer-term or more intensive program services. In addition, federal officials noticed that local program administrators were failing to terminate (process the exit) of enrollees who, while no longer receiving services, were unemployed. By holding back poorly performing enrollees even when these enrollees were no longer in contact with the program, training centers could boost their performance scores. Department of Labor officials subsequently closed this loophole by limiting the time an idle enrollee could stay in the program to 90 days (see Barnow and Smith, 2002; Courty and Marschke, 2007a; Courty et al., 2008).

In effect, the experience of real world incentive designers in both the public and private sectors offers evidence that the nature of a performance measure's distortions is frequently unknown to the incentive designer before implementation. In the public sector the challenges may be greater because the incentive designer may be far removed from the front line worker she is trying to manage and may have less detailed knowledge of the mechanics of the position and/or the technology of production. In the case of JTPA, Congress outlined in the Act the performance measures for JTPA managers, and staff at the U.S. Department of Labor refined them. It is unlikely that these parties had complete, first-hand knowledge of job training processes necessary to fully anticipate the distortions in the performance measures that they specified.

The static model adapted to a dynamic context

Courty and Marschke (2003, 2007b) adapt the standard principal-agent model to show how organizations manage performance measures when gaming is revealed over time. Their model assumes the principal has several imperfect performance measures to choose from, each capturing an aspect of the agent's productive effort, but each also is distorted. The measures have their own distinct weaknesses, and these weaknesses are unknown to the incentive designer ex ante. The agent, however, knows precisely how to exploit each weakness, because of his day-to-day familiarity with the technology of production.

Courty and Marschke (2007b) assume the agent exerts effort on a set of projects or tasks $\alpha \in A$, which is exogenously given. The index α captures the production environment, which varies from project to project. In the context of JTPA, for example, one might view as a project a group of persons to be trained who share the same particular demographic characteristics. The agent privately observes each project's type α after signing the contract but before making her effort decisions. The principal does not observe α or effort. The agent's actions on each project influence two performance measures as well as the value of the program or organization. Assuming as in (1) and (2) organizational value and performance are linear in **a**, the value and performance outcomes generated for project α are

$$V_\alpha(\mathbf{a}) = f_{0,\alpha}a_0 + f_{1,\alpha}a_1 + f_{2,\alpha}a_2$$

$$P_{1,\alpha}(\mathbf{a}) = g_{0,\alpha}a_0 + g_{1,\alpha}a_1 + w_{1,\alpha}a^w{}_1$$

$$P_{2,\alpha}(\mathbf{a}) = g_{0,\alpha}a_0 + g_{2,\alpha}a_2 + w_{2,\alpha}a^w{}_2 \qquad (3)$$

where $g_{i,\alpha}=f_{i,\alpha}+\eta_{i,\alpha}$. Because the model focuses on the distortions in the performance measures, it ignores the additive noise terms. a_0 captures the dimension of effort that is common to both performance measures and to V. In addition, both measures imperfectly capture different dimensions of effort (a_1, a_2) and both have (different) gaming dimensions $w_{1,\alpha}a^w{}_1$ and $w_{1,\alpha}a^w{}_1$. A performance measure is distorted if it either imperfectly captures the marginal productivity of effort ($\eta_{i,\alpha} \neq 0$) for some α, or if the agent can take actions that increase the performance measure but not value-added, i.e. $w_{i,\alpha} > 0$ for some α.

The performance outcome for measure i is the sum of performance outcomes over all projects, $P_i = \Sigma_\alpha P_{i,\alpha}(\mathbf{a})$. Courty and Marschke (2003) assume multiple periods, and in any one period, the incentive designer is restricted to use only one measure. In the first period, the incentive designer chooses a performance measure at random (as the attributes of the measure are unknown ex ante) and a contract (s, b_p). The model assumes the agent chooses her effort to maximize her compensation in the current period net of effort costs (i.e., myopically; see below). The incentive designer monitors the agent's actions and eventually learns the extent of the measure's distortion. If the performance measure proves to be less distorted than expected, the incentive designer increases the weight on the measure in the next period, and if the distortion is greater than expected, she replaces it with the other measure. The model implies that major changes in performance measures are more likely to take place early in the implementation of performance measurement systems, before the incentive designer acquires knowledge of the feasible gaming strategies.

After the learning comes to an end, a dynamic learning model such as this delivers the same implications for the relations among incentive weights, alignment, risk aversion, and so on as does the static model. A new implication, however, is that the alignment between a performance measure and the true goal decreases as a performance measure is activated, or as it is more heavily rewarded, and increases as the measure is retired. This occurs because once the measure is activated, an agent will turn his focus on it as an objective and begin to explore *all* strategies for raising it—not just the ones that also improve organizational value.

For example, before public schools attached high stakes to achievement test outcomes, it indeed may have been the case that student test scores were strongly positively correlated with their educational achievement levels. Once educators came under strong pressures to increase student proficiency levels, increasing test scores rather than increasing student learning became the focal objective, and some educators began to look for quicker and easier ways to boost student performance. Some of these ways, such as teaching to the test, were effective in raising student test scores, but at the expense of students' learning in other academic areas. Thus, we expect that the correlation between student test scores and learning likely fell with the introduction of sanctions for inadequate yearly progress on these performance measures, generally reducing the value of test scores for assessing student learning and school performance.

The problem with using correlation (and other measures of association) to choose performance measures is that the correlation between an *inactivated* performance measure and value added does not pick up the distortion inherent in the performance measure. The change in the correlation structure before and after activation, as Courty and Marschke (2007b) show, reflects the amount of distortion in a performance measure and thus can be used to measure the distortion in a performance measure. They formally demonstrate that a measure i is distorted if and only if the regression coefficient of V_α on $P_{i,\alpha}$ decreases after the measure's activation,[8] and then using data from JTPA, they find that the estimated regression coefficients for some measures indeed fall with their introduction, suggesting the existence of performance measure distortion in JTPA. These findings corroborate their previous work (Courty and Marschke, 2004) that used the specific rules of the performance measurement system to demonstrate the existence of distortions in JTPA performance measures.

This dynamic model shows that common methods used to select performance measures have important limitations. As Baker (2002) points out, researchers and practitioners frequently evaluate the usefulness of a performance measure based on how well the measure predicts the true organizational objective, using statistical correlation or other measures of association (see, for example, Ittner and Larcker, 1998; Banker, Potter, and Srinivasan, 2000; van Praag and Cools, 2001; Burghart et al., 2001; and Heckman, Heinrich, and Smith, 2002). As discussed earlier, Baker makes the point that in practice, one has to know how much these correlations are capable of revealing about the agents' choices; if one assumes the linear formulation of noisy performance measures and value, as in (1) and (2), a correlation of the two will pick up only the correlation in the noise terms (ε,φ), which tells one nothing about the agent's choices. An additional implication of the dynamic model outlined above (with random marginal

[8] They also show that measure i is distorted if the correlation of $P_{i,\alpha}$, and V_α decreases with activation These results also hold in the case where both measures are used simultaneously, and the relative weight on measure i increases.

productivities as in (3)) is that the association between a performance measure and organizational value is endogenous, casting further doubt on these methods for validating performance measures.

Learning in a dynamic context

In the dynamic model described above, the assumption of multiple periods introduces the possibility that agents will behave strategically. The model could be adapted to permit the agent to anticipate that the principal will replace performance measures once she learns that the agent is gaming. The main predictions of the model would still hold, however, as it would still be optimal for the agent to game.

The multi-tasking model could also be extended to permit learning by the agent. Koretz and Hamilton (2006) describe evidence that the difference between states' and school districts' performance on standardized tests and student learning appears to grow over time. Performance inflation is interpreted as evidence that agents learn over time how to manipulate test scores. Thus, just as the static model could be reformulated to allow for the principal learning about how effective a performance measure is, it could be reformulated to allow the agent to learn how to control the performance measure. In such circumstances, little gaming would take place early on, but gaming would increase as the agent acquired experience and learned the measure-specific gaming technology. The amount of gaming that takes place in a given period, therefore, will depend on how distorted the measure was to start with, how long the agent has had to learn to game the measure, and the rate of learning. In other words, although a performance measure may be very successful when it is first used, its effectiveness may decline over time, and it may be optimal at some point to replace it.

We use the example of the Wisconsin Works (W-2) welfare-to-work program, in which the state modified its performance-based contracts with private W-2 providers over successive

(2-year) contract periods, to illustrate learning on the part of both the principal and agents (see Heinrich and Choi, 2007). In the first W-2 contract period (beginning in 1997), performance measures were primarily process-oriented (e.g., the client-staff ratio, percent of clients with an employability plan, etc.), with the exception of a single outcome measure (the percent of W-2 cases that returned to a W-2 grant after being placed in a job). Provider profits in this period were tied primarily to low service provision costs, and welfare caseload declines that were already in progress made it easy for contractors to meet the performance requirements with little effort and expense, particularly if they discouraged participation of the harder-to-serve. W-2 providers consequently realized unexpectedly large profits (from the state's perspective).

The state subsequently responded in the second period by establishing five new outcome measures intended to increase contractors' focus on service *quality*—employment rate, average wage rate, two job retention measures, and employer health insurance benefits—and reduced the process measures to just two. Although profit levels were also restricted in this period, W-2 providers continued to make sizeable profits, partly because some performance targets were set too low to be meaningful, but also because contractors identified new ways to adjust client intake and activities to influence the new performance outcome levels.

The third contract period represented a significant turning point, with additional changes that reflected the state's cumulative learning over the first two periods. The state defined four categories of performance measures, adding new measures of quality of case management services, customer satisfaction, and financial management (among others), and assigning different weights to them. In addition, two levels of bonuses were defined in the contracts (bonuses with restrictions and without restrictions, the latter for higher levels of performance achievement), and the "priority outcomes" performance measures were accorded two-thirds of the weight in the bonus calculations. Again, the W-2 contractors appeared to grasp the new incentives and responded accordingly. For example, in the second contract period when the educational attainment performance measure was *optional*, only three of 71 agencies (4%) met the base performance level. After this same measure was *required* and assigned weight in the determination of bonuses in the third period, 59 (88%) of W-2 agencies met the minimum (base) performance level. Alternatively, the state attempted to use an average wage/earnings change measure in the second and third contract periods, but it changed this performance measure from "required" to "information-only" midway through the third contract period due to concerns about data accuracy. As a result (after being made optional), *measured* performance on this standard declined dramatically between these two contract periods, from 100% of agencies meeting it to just 42% of agencies achieving base performance levels. Statistical analyses of UI wage record data confirmed, however, that there were no *real* (statistically significant) differences in participant average wage changes between these two periods. Rather, it appears that agents were probably shifting their gaming efforts, such as changing participants' assignments to different "tiers" of W-2 participation that affected their inclusion in the performance calculations, according to the weights placed on the different standards. In fact, the state decided to retire this and other measures in the fourth contract period and also rescinded the bonus monies that had been allocated to reward performance.

This example suggests that if an agent is learning along with the principal, the dynamics may get very complicated. And if performance measures degrade over time, then there will be a dynamic to measurement systems that will not necessarily stop. As discussed earlier, this type of dynamic also appears to be present in the selective enrollment of clients into the JTPA and WIA programs. In order to discourage enrollment practices strategically designed to increase performance (regardless of program value added), policymakers first developed separate performance measures for the "hard-to-serve" in the JTPA program. However, states and localities could choose to place more or less weight on these measures, and front-line workers learned over time how to select the most able from among the classes of hard-to-serve defined by the measures (e.g., the most employable among welfare recipients). The WIA program introduced universal access in part to ameliorate these selection problems, yet access to more intensive training is still limited by a sequential enrollment process. Thus, WIA also added an earnings change performance measure intended to discourage the favorable treatment of clients with stronger pre-program labor force histories. After only four years, however, the U.S. Department of Labor recently concluded that it was encouraging a different type of strategic enrollment, i.e., discouraging service to workers with higher earnings at their last job in order to achieve a higher (earnings change) performance level. A GAO (2002:14) report confirmed in interviews with WIA program administrators in 50 states the persistence of this type of behavioral response: “the need to meet performance levels may be the driving factor in deciding who receives WIA-funded services at the local level.”

Setting and adjusting performance standards

Another open dynamic issue concerns the growing number of performance measurement systems that have built requirements for “continuous performance improvement” or “stretch targets” into their incentive systems. We may perhaps observe this because principals believe that the optimal effort changes over time, as agents learn and improve in their use of the production technology, and that this learning in turn increases the return at the margin of supplying effort. However, if agents’ efforts influence the rate at which performance improvements are expected, this may contribute to an unintended ratchet problem. In effect, if the agent learns or adjusts more quickly than the principal anticipates, this may lead to under-provision of effort.

One way that the principal can eliminate the ratchet effect is by committing to holding the performance target fixed, or more realistically, by committing to strict rules for changing the standard. The idea is that such a commitment will eliminate apprehension about the ratchet effect and reinforce incentives for effort. In the WIA program, states established performance standards for three-year periods, using past performance data to set targets for the first year and building in predetermined expectations for performance improvements in the subsequent two years. Unfortunately, over the first three years of the program (July 2000-03), an economic recession made it significantly more challenging for states to meet their performance targets. As performance standard adjustments for economic conditions are not standard practice in WIA, the unadjusted performance expectations were in effect higher than anticipated, which triggered a number of gaming responses intended to mitigate the unavoidable decline in performance relative to the standards (see Courty et al., 2005).

The use of “stretch” or benchmarked targets (i.e., typically based on past performance assessments) has been criticized for its lack of attention to or analysis of the process or capability for achieving continuous performance improvements. Castellano et al. (2004) identify the failure to understand variation in process or the inherent inability of a stable process to regularly achieve point-specific targets as a fatal flaw of performance measurement systems. In effect, if performance measurement is to produce accurate knowledge of the value-added of government activities, then it is essential to understand both the relationship of government activity to performance outcomes and the factors that influence outcomes but cannot (or should not) be controlled by public managers. That is, estimates of performance are more likely to accurately (and usefully) reflect the contribution (or value-added) of public managers and program activities to any changes in performance if performance expectations are adjusted for factors that are not controlled in production.

Yet in a recent review of the use of performance standard adjustments in government programs, Barnow and Heinrich (forthcoming) show that their use is still relatively rare, and that concerns that adjustments are likely to be noisy, biased and/or unreliable are prevalent. They also suggest, however, that such criticisms are likely overstated relative to the potential benefits of obtaining more useful information for improving program management and discouraging "gaming" responses (such as altering who is served and how to influence measured performance).[9] The Job Corps program, for example, uses regression models to adjust performance standards for five of 11 performance measures that are weighted and aggregated to provide an overall report card grade (U.S. Department of Labor, 2001). The stated rationale is that "by setting individualized goals that adjust for differences in key factors that are beyond the operator's control, this helps to 'level the playing field' in assessing performance" (U.S. Department of Labor, 2001, Appendix 501, p. 6). Some of the adjustment factors include: participant age, reading and numeracy scores, occupational group, and state and county economic characteristics. In practice, the Job Corps adjustment models frequently lead to large differences in performance standards across the centers. In 2007, the diploma/GED attainment rate standard ranged from 40.5 percent to 60.8 percent, and the average graduate hourly wage standard ranged from $5.83 per hour to $10.19 per hour. Job Corps staff report that center operators find the performance standards adjustments to be credible and fair in helping to account for varying circumstances across centers.

In fact, three different studies of the shift away from formal performance standard adjustment procedures under WIA concluded that the abandonment of the regression model adjustment process (used under JTPA) led to standards viewed as arbitrary, increased risk for program managers, greater incentives for creaming among potential participants, and other

[9] Courty, Kim, and Marschke (2008) show that the use of regression-based models in JTPA indeed seems to have influenced who in JTPA was served.

undesirable post-hoc activities to improve measured performance (Heinrich, 2004; Social Policy Research Associates, 2004; Barnow and King, 2005). Barnow and Heinrich (forthcoming) conclude that more experimentation with performance adjustments in public programs is needed, or we will continue to be limited in our ability to understand not only whether they have the potential to improve the accuracy of performance assessments, but also if they contribute to improved performance over time as public managers receive more useful feedback about their programs' achievements (or failures) and what contributes to them.

Toward better systems for measuring performance improvement

The fact that a majority of public sector performance measurement systems adopt a linear or "straight-line" approach to measuring performance improvement or are nonlinear in design also contributes importantly to the implementation problems described above. A straight-line approach typically defines a required (linear) rate of performance improvement from an initial score or target level and may also specify an ending value corresponding to a maximum performance *level* (e.g., 100% proficiency). Alternatively, in nonlinear performance incentive system designs, performance below a specified numerical standard receives no recognition or reward, and/or performance at or above the standard is rewarded with a fixed-size bonus.

A common flaw of straight-line performance improvement models for establishing performance expectations, explain Koretz and Hamilton (2006), is that they are "rarely based on empirical information or other evidence regarding what expectations would be realistic" (p. 540). Although this is also a concern in nonlinear incentive systems, the latter induce other distortions that are potentially more problematic. In their illustration of the principle of equal compensation in incentives, Holmstrom and Milgrom (1991) predicted that the implementation of nonlinear incentive contracts would cause agents to reallocate effort away from tasks or projects (or clients/subjects) for which greater effort is required to push the project or client over the threshold, or for which the project or client would exceed the threshold on their own (see also Holmstrom and Milgrom, 1987). In other words, agents would exert more effort in activities or with clients that were likely to influence performance close to the standard.

For example, in education systems with performance assessments based on specific standards for proficiency, we are more likely to observe a reallocation of effort to "bubble kids," i.e., those on the cusp of reaching the standard (Pedulla et al., 2003; Koretz and Hamilton, 2006). Figlio and Rouse (2005) found that Florida schools that faced pressure to improve the fraction of students scoring above minimum levels on comprehensive assessment tests focused their attention on students in the lower portion of the achievement distribution and on the subjects/grades tested on the exams, at the expense of higher-achieving students and subjects not included in the test. In their study of school accountability systems in Chicago, Neal and Schanzenbach (2007) similarly concluded that a focus on a single proficiency standard would not bring about improvement in effort or instruction for *all* students, likely leaving out those both far below and far above the standard. The WIA example discussed earlier shows another form of this response. In the face of worsening economic conditions and the requirement to demonstrate performance improvements on an earnings change measure, case workers restricted access to services for dislocated workers for whom it would be more difficult to achieve the targeted increase in earnings (i.e., above their pre-program earnings levels).

The high-stakes elements of recent public sector performance incentive systems—such as the threat of school closure under the No Child Left Behind Act and the potential to win or lose funding or to be reorganized in public employment and training programs—have exacerbated these gaming responses and contributed to a growing interest in the use of value-added approaches in performance measurement. In general, value-added models aim to measure productivity *growth* or improvement from one time point to another and the process by which any such growth or gains are produced, taking into account initial productivity or ability levels and adjusting for factors outside the control of administrators (e.g., student demographics in schools). Neal and Schanzenbach (2007) show that a value-added system with performance measured as the total net improvement in school achievement has the potential to equalize the marginal effort cost of increasing a given student's test score, irrespective of the student's initial ability. In effect, a value-added design eliminates (more than any alternative model) the incentive to strategically focus effort on particular subgroups of public program clients to improve measured performance.

Neal and Schanzenbach also acknowledge, however, that taking into account limits of or constraints on the technology of production, and in the case of public schools, students' initial distribution of ability, there may still be circumstances in which little effort would be exerted to help some student subgroups improve. Indeed, interest in implementing value-added performance measurement systems in public schools is expanding largely because they have the potential to produce useful information for policymakers in understanding how school/policy inputs and interventions relate to and can be better targeted to achieve net improvements in student outcomes (Roderick, Jacob and Bryk, 2002, 2004; Thorn and Meyer, 2006). Described in terms of the value and performance equations discussed above (in 4), it is a goal of the value-added approach to better understand and motivate the dimension of effort (a_0) that is common to both performance measures (P) and to value (V). Thorn and Meyer of the Value-Added Research Center at the Wisconsin Center for Education Research, for example, report how school district staff, with the support of researchers, are using value-added systems to not only measure teacher and classroom performance, but to also diagnose problems, track student achievement over time and assess the decay effects of interventions, and to remove incentives to teach narrowly to test outcomes or focus on short-term gains. Of course, value-added performance measurement systems are still in the early stages of development and implementation in the public sector, and as in other incentive system designs, researchers are already identifying emerging measurement challenges and complex dynamics of which the implications will only become better known with time, experience and careful study of their use (Harris and Sass, 2006; Lockwood et al, 2007; Koedel and Betts, 2008; Rothstein, J., 2008).

In fact, the development and expanding use of value-added systems for measuring performance in education appears to be encouraging exactly the type of "interactive dialogue" that Moynihan proposes to support more effective use of performance information and organizational learning. In February 2008, Vanderbilt University's National Center on Performance Incentives convened a conference that brought together federal, state and local education policymakers, teachers and school support staff, and academics and school-based researchers to debate the general issues of incentives for improving educational performance as well as the evidence in support of approaches such as value-added performance measurement systems and other strategies for assessing school performance. The conference organizers also invited participants who could illuminate lessons learned about performance measurement system design and implementation in other sectors, such as health care, municipal services, and employment and training (Rothstein, R., 2008). Two months later (April, 2008) the Wisconsin Center for Education Research at the University of Wisconsin-Madison organized a conference to feature and discuss the latest research on developments in value-added performance measurement systems, which was likewise attended by both academic researchers and school district staff engaged in implementing and using value-added systems. As advocated by Moynihan (p. 207), these efforts appear to be encouraging the use value-added performance measurement systems as a "practical tool" for performance management and improvement, with "more realistic expectations" and "a clearer map of the difficulties and opportunities in implementation."

CONCLUSIONS

Drawing from examples of performance measurement systems in public and private organizations, we have sought to highlight key challenges in the design and implementation of performance measurement systems in the public sector. The complicated nature and technology of public programs compels performance measurement schemes to accommodate multiple and difficult-to-measure goals. One lesson we draw from the literature is that incentive schemes will be more effective if they are not simply grafted onto the structure of an agency, regardless of its complexities or the heterogeneity of employees and work tasks. For example, an incentive system designer in a multitask environment, where some tasks are measurable and others are not, may be able to develop a more effective performance incentive scheme if care is taken to understand what motivates employees and to assign or reallocate tasks across workers accordingly. Assigning work so that one group of agents performs only measurable tasks and placing another group of intrinsically motivated workers in positions where performance is difficult to measure would exploit the motivating power of incentives for some workers and attenuate the moral hazard costs from lack of incentives for the others. The usefulness of this strategy depends, of course, on the ability to identify intrinsically motivated workers and to facilitate a structural or functional separation of work tasks, which may be more or less feasible in some public sector settings.

Performance incentive system designers also need to appreciate and confront the evolutionary dynamic that appears to be an important feature of many performance measurement schemes. As illustrated, an incentive designers' understanding of the nature of a performance measure's distortions and employees' means for influencing performance is typically imperfect prior to implementation. It is only as performance measures are tried, evaluated, modified and/or discarded that agents' responses become known. This type of performance measurement monitoring to assess a measure's effectiveness and distortion requires a considerable investment on the part of the principal, one that public sector program managers have neglected or underestimated in the past (Hefetz and Warner, 2004; Heinrich and Choi, 2007).

Of course, a complicating factor is that performance measures can be gamed. Agents (or employees), through their day-to-day experience with the technology of production, come to know the distinct weaknesses or distortions of performance measures and how they can exploit them. If it takes the agent time to learn how to game a new performance measure, an equilibrium solution may be a periodic revision of performance measures, or a reassignment of agents. However, if both principals and agents are learning over time, the dynamic is likely to become more complex. And depending on the relative speeds of principal and agent learning and the extent to which performance measures degrade over time, this dynamic to measurement systems will not necessarily end. If the principal learns faster than the agent, the usefulness of a performance measure is more likely to increase, but if the agent learns faster how to manipulate a measure, its usefulness will decline and the measure may ultimately be discarded.

Furthermore, in a model with bilateral learning across multiple periods, not only can the form of the performance incentives have this never-ending dynamic, but the welfare path can also follow a very complicated course. For example, if agent learning about gaming is significant, one might realize an immediate improvement in social welfare with the implementation of performance incentives, followed by a drop in welfare as the agent learns how to game. If principal learning is important, an initial drop in welfare with the implementation of performance incentives may be followed by an improvement as the principal learns how to make adjustments that increase the efficiency of the system. And with a combination of principal learning and agent learning, numerous paths in welfare will be possible. This has important implications not only for the principal's decision about whether to implement an incentive system, but also for our ability to evaluate the effectiveness of incentive schemes. That is, whether or not we conclude the incentive system is successful will likely depend on where in this dynamic cycle we are evaluating the program. We suggest that additional empirical and theoretical exploration of the dynamic aspects of performance measurement systems could be fruitful for generating additional insights and policy recommendations.

In practice, faced with these challenges, some incentive designers will “quit the game,” as in the recent case of the Temporary Assistance for Needy Families high performance bonus system. However, there are promising examples of possible payoffs to persistence. The State of Wisconsin, in its fourth W-2 contract period, also appeared to give up on its performance incentive system, retracting its performance measures and funding allocations for bonuses. However, in the latest contract period, it has made a comeback with the undertaking of a significant structural and management reorganization and new provisions for performance-based contracting. The state shifted to a four-year (2006-2009) contract to “promote stability,” more efficiently use administrative funds, and strengthen partnerships with other private service providers. In addition, it adopted new contract incentives that require larger W-2 contractors to “earn” 20 percent of their administrative and service contract funds through the achievement of specific outcomes; that is, reimbursement of costs is conditional on these program outcomes being met. Perhaps most significant, the state broke its larger contracts and is separately contracting out for W-2 case management, job development and placement, and Supplemental Security Income advocacy activities to allow for greater specialization, competition in service delivery, improved targeting of services to clients, and performance measures and incentives more closely aligned with service objectives. As with value-added performance measurement systems in education, only time will tell if the incentive designers have uncovered the right incentive scheme, although experiences to date in public employment and training programs suggest that the dynamic is unlikely to end in the current round.

REFERENCES

Akerlof, G. and R. Kranton (2005): "Identity and the Economics of Organizations", Journal of Economic Perspectives, 19(1), Winter.

Asch, B. J. (2005). The Economic Complexities of Incentive Reforms. In Klitgaard, R. & and P. C. Light (Eds.), High-Performance Government: Structure, Leadership, Incentives (pp. 309-342). Santa Monica, CA: RAND Corporation.

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

Baker, George P. (1992). "Incentive Contracts and Performance Measurement." *Journal of Political Economy*, 100, pp. 598-614.

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

Banker, R., Potter, G., and Srinivasan, D. (2000). "An Empirical Investigation of an Incentive Plan that Includes Nonfinancial Performance Measures," *The Accounting Review*, 75(1), pp. 21-39.

Barnow, Burt A. and Carolyn J. Heinrich. Forthcoming. "One Standard Fits All? The Pros and Cons of Performance Standard Adjustments." *Public Administration Review.*

Barnow, Burt S. and Christopher T. King. (2005). *The Workforce Investment Act in Eight States.* Albany, NY: The Nelson A. Rockefeller Institute of Government.

Barnow, Burt and Smith, Jeffery. (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.

Bertelli, Anthony M. and Lynn, Laurence E., Jr., (2006). *Madison's Managers: Public Administration and the Constitution.* Baltimore: The Johns Hopkins University Press.

Besley, Timothy and Ghatak, Maitreesh. (2003). "Incentives, Choice and Accountability in the Provision of Public Services." London School of Economics, Institute for Fiscal Studies, Working paper 03/08.

Bevan, G. & Hood, C. (2006). What's Measured is What Matters: Targets and Gaming in the English Public Health Care System. Public Administration, 84(3), 517-538.

Bloom, Howard S., Carolyn J. Hill, and James Riccio. (2005). "Modeling Cross-Site Experimental Differences to Find Out Why Program Effectiveness Varies." In Howard S. Bloom (ed). *Learning More from Social Experiments: Evolving Analytic Approaches.* (New York: Russell Sage Foundation), pp. 37-74.

Brudney, Jeffrey L., Sergio Fernandez, Jay Eungha Ryu and Deil S. Wright. (2005). “Exploring and Explaining Contracting Out: Patterns among the American States**.”** *Journal of Public Administration Research and Theory* 15(3): 393-419**.**

Castellano, Joseph F., Young, Saul and Harper A. Roehm. (2004). “The Seven Fatal Flaws of Performance Measurement. *The CPA Journal* Vol. LXXIV, No. 6: 32-35.

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

Courty, Pascal and Marschke, Gerald. (2003b). “Dynamics of Performance-Measurement Systems,” *Oxford Review of Economic Policy*, 19(2), pp. 268-284.

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

Courty, P. and Marschke, G. (2007a). "Making Government Accountable: Lessons from a Federal Job Training Program," *Public Administration Review*, forthcoming.

Courty, P. and Marschke, G. (2007b). "A General Test for Distortions in Performance Measures," *The Review of Economics and Statistics*, forthcoming.

Courty, Pascal, Heinrich, Carolyn J. and Gerald R. Marschke. (2005). “Setting the Standard in Performance Measurement Systems.” *International Public Management Journal,* 8(3): 321-347.

Courty, Pascal, Kim, Do Han, and Gerald R. Marschke. (2008). “Curbing cream-skimming: Evidence on enrollment incentives.” Working paper.

Davis, J.H., L. Donaldson & Schoorman, F.D. (1997). Toward a Stewardship Theory of Management. Academy of Management Review, 22(1), 20-47.

Dias, J. J. and S. Maynard-Moody (2007). “For-Profit Welfare: Contracts, Conflicts, and the Performance Paradox.” *Journal of Public Administration Research and Theory* 17(2):189-211.

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

Ewalt, JoAnn and Edward Jennings. (2004). “Administration, Governance, and Policy Tools in Welfare Policy Implementation.” *Public Administration Review* 64(4**):** 449 – 462.

Figlio, David N. and Cecilia E. Rouse. (2005). “Do Accountability and Voucher Threats Improve Low-performing Schools?” National Bureau of Economic Research Working Paper 11597.

Francois, P. and M. Vlassopoulos. (2008). **“**Pro-social Motivation and the Delivery of Social Serviceshttp://cesifo.oxfordjournals.org/cgi/content/full/ifn002v1”, CESifo Economic Studies.

Frederickson, David G. and George H. Frederickson. (2006). *Measuring the Performance of the Hollow State*. Washington, DC: Georgetown University Press.

Frey, B. S. and M. Benz. (2005). "Can Private Learn From Public Governance?" *The Economic Journal*, 115: F377-F396.

Harris, Douglas and Tim R. Sass. (2006). Value-Added Models and the Measurement of Teacher Quality. Unpublished manuscript, Department of Economics, Florida State University, Tallahassee.

Hatry, Harry P. (1999) *Performance Measurement: Getting Results*. Washington, D.C.: The Urban Institute Press.

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

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

Heckman, James J., Carolyn J. Heinrich and Jeffrey A. Smith. (1997). "Assessing the Performance of Performance Standards in Public Bureaucracies." *American Economic Review*, Vol. 87, No. 2: 389-395.

Hefetz, A. and Warner, M. (2004). "Privatization and Its Reverse: Explaining the Dynamics of the Government Contracting Process." *Journal of Public Administration Research and Theory* 14(2): 171-190.

Heinrich, Carolyn J. (2007). "False or Fitting Recognition? The Use of High Performance Bonuses in Motivating Organizational Achievements." *Journal of Policy Analysis and Management* 26(2): 281-304.

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

Heinrich, Carolyn J. (1995). "Public Policy and Methodological Issues in the Design and Evaluation of Employment and Training Programs at the Service Delivery Area Level." Ph.D. dissertation, University of Chicago.

Heinrich, Carolyn J. and Youseok Choi. (2007). "Performance-based Contracting in Social Welfare Programs." *The American Review of Public Administration* 37(4): 409-435.

Heinrich, Carolyn J. and Elizabeth Fournier. (2005). "Instruments of Policy and Administration for Improving Substance Abuse Treatment Practice and Program Outcomes." *Journal of Drug Issues,* 35(3): 481-500.

Hill, Carolyn J. (2006). “Casework Job Design and Client Outcomes in Welfare-to-Work Offices.” *Journal of Public Administration Research and Theory* 16(2): 263-288.

Holmstrom, Bengt. (1982). “Moral Hazard in Teams.” *Bell Journal of Economics* 13: 324-40.

Holmstrom, Bengt and Paul Milgrom. (1991). “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” *Journal of Law, Economics and Organization* 7 (January): 24-52.

Holmstrom, Bengt and Paul Milgrom. (1987). “Aggregation and Linearity in the Provision of Intertemporal Incentives.” Econometrica 55: 303-328.

The Hoover Commission Report. (1949). New York: McGraw-Hill Book Company, Inc.

Houston**,** David J. (2006). "’Walking the Walk’ of Public Service Motivation: Public Employees and Charitable Gifts of Time, Blood, and Money.” *Journal of Public Administration Research and Theory* 2006 16(1):67-86.

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

Jacob, B. A. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago." *Journal of Public Economics* 89(5-6): 761-796.

Jacob, Brian A., and Levitt, Steven D. (2003). “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating.” *Quarterly Journal of Economics*, 118(3), pp. 843-877.

Koedel, C. and J. Betts. (2008). “Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation.” National Center on Performance Incentives Working paper 2008-21, Vanderbilt University, Peabody College.

Koretz, D., and Hamilton, L. S. (2006). “Testing for accountability in K-12.” In R. L. Brennan (Ed.), *Educational measurement* (4th ed.), pp. 531-578. Westport, CT: American Council on Education/Praeger.

Lockwood, J.R., Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le and Jose Felipe Martinez. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement,* 44:47-67.

Lynn, Laurence E., Jr., Heinrich, Carolyn J., and Hill, Carolyn J. (2001) *Improving Governance: A New Logic for Research.* Washington, DC: Georgetown University Press.

May, Peter J. and Soren Winter. Forthcoming. “Politicians, Managers, and Street-Level Bureaucrats: Influences on Policy Implementation.” *Journal of Public Administration Research and Theory.*

Moe, Ronald C. (1982). “A New Hoover Commission: A Timely Idea or Misdirected Nostalgia?” *Public Administration Review*, 42 (3): 270-277.

Moynihan, D. P. (2008). *The Dynamics of Performance Management: Constructing Information and Reform.* Washington D.C.: Georgetown University Press.
Murdock, Kevin. 2002. “Intrinsic Motivation and Optimal Incentive Contracts.” *The RAND Journal of Economics* 33(4): 650-671.

Murphy, K. R. and Cleveland, J. N. (1995). *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives.* Thousand Oaks: Sage Publications.

Neal, Derek and Diane Schanzenbach. (2007). “Left Behind by Design: Proficiency Counts and Test-Based Accountability.” Working paper, University of Chicago.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K. Ramos, M. A., and Miao, J. (2003). *Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers.* Boston: National Board on Educational Testing and Public Policy.

Perry, J. L. & Porter, L. (1982). Factors Affecting the Context for Motivation in Public Organizations. Academy of Management Review, 7(1), 89-98.

Perry, J. L. & Wise, L. R. (1990). The Motivational Bases of Public Service. Public Administration Review, 50 (May/June): 367-373.

Prendergast, Canice. (2007). “The Motivation and Bias of Bureaucrats.” *The American Economic Review* 97(1): 180-196.

Radin, B. (2006). *Challenging the Performance Movement.* Washington, D.C.: Georgetown University Press.

Rainey, Hal G. (2006). "Reform Trends at the Federal Level with Implications for the States: The Pursuit of Flexibility and the Human Capital Movement." In J. Edward Kellough and Lloyd G. Nigro, (Eds.), *Civil Service Reform in the States: Personnel Policies and Politics at the Sub-National Level.* Albany: State University of New York Press, pp: 33-58.

Rainey, Hal G. (1983). "Public Agencies and Private Firms: Incentive Structures, Goals, and Individual Roles," *Administration and Society*, Vol. 15, No. 2: 207-242.

Roderick, Melissa, Brian A. Jacob, and Anthony S. Bryk. (2002). "The Impact of High-Stakes Testing in Chicago on Student Achievement in the Promotional Gate Grades." *Educational Evaluation and Policy Analysis* 24(4): 333-357.

Roderick, Melissa, Brian A. Jacob, and Anthony S. Bryk. (2004). “Summer in the City: Achievement Gains in Chicago’s Summer Bridge Program." *Summer Learning: Research, Policies, and Programs.* Ed. Geoffrey D. Borman and Matthew Boulay. Lawrence Erlbaum Associates.

Rothstein, J. (2008). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. Unpublished Manuscript, Princeton University.

Rothstein, R. (2008). Holding Accountability to Account: How Scholarship and Experience in Other Fields Inform Exploration of Performance Incentives in Education. National Center on Performance Incentives Working paper 2008-04, Vanderbilt University, Peabody College.

Scott, W.D., Clothier, R.C. and Spriegel, W.R. (1941). *Personnel Management*. New York: McGraw-Hill.

Social Policy Research Associates. (2004). *The Workforce Investment Act after Five Years: Results from the National Evaluation of the Implementation of WIA.* Oakland, CA: Social Policy Research Associates.

Stone, D. (1997). *Policy paradox: The art of policy decision making*. New York, NY: WW Norton & Company.

Taylor, F., (1911). “Principles and Methods of Scientific Management.” *Journal of Accountancy,* 12/2: 117-24.

Thompson, James D. (1967) *Organizations in Action*. New York: McGraw-Hill Book Company.

Thorn, C. A. and Meyer, R. H. (2006). Longitudinal data systems to support data-informed decision making: A tri-state partnership between Michigan, Minnesota and Wisconsin. Report of the Wisconsin Center for Education Research.

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.

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

Van Slyke**,** David M. (2007). “Agents or Stewards: Using Theory to Understand the Government-Nonprofit Social Service Contracting Relationship.” *Journal of Public Administration Research and Theory* 2007 17(2):157-187.

Walker, R M. & Boyne, G. A. (2006). Public Management Reform and Organizational Performance: An Empirical Assessment of the U.K. Labour Government’s Public Service Improvement Strategy. Journal of Policy Analysis and Management, 25(2), 371-393.

Wilson, W. (1887). The Study of Administration. Political Science Quarterly, 2(2), 197-222.