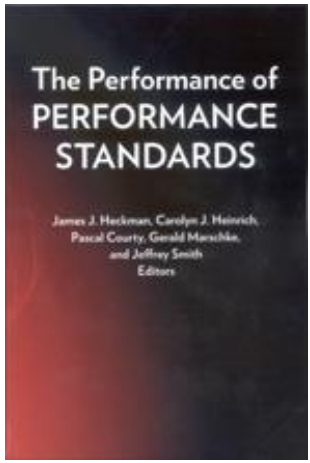
W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH

Upjohn Institute Press

Measuring Government Performance: An Overview of Dysfunctional Responses



Pascal Courty
University of Victoria

Gerald Marschke
University at Albany, SUNY

Chapter 7 (pp. 203-230) in:
The Performance of Performance Standards
James J. Heckman, Carolyn J. Heinrich, Pascal Courty, Gerald Marschke, and Jeffrey Smith, eds.
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 2011
DOI: 10.17848/9780880993982.ch7

The Performance of Performance Standards

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

2011

W.E. Upjohn Institute for Employment Research
Kalamazoo, Michigan

Library of Congress Cataloging-in-Publication Data

The performance of performance standards / James J. Heckman . . . [et al.], editors.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-88099-292-3 (pbk. : alk. paper)
ISBN-10: 0-88099-292-1 (pbk. : alk. paper)
ISBN-13: 978-0-88099-294-7 (hardcover : alk. paper)
ISBN-10: 0-88099-294-8 (hardcover : alk. paper)
1. Government productivity. 2. Performance standards. 3. Civil service—Personnel management. I. Heckman, James J. (James Joseph)
JF1525.P67P476 2011
352.6'7—dc22

2011007877

W.E. Upjohn Institute for Employment Research
300 S. Westnedge Avenue
Kalamazoo, Michigan 49007-4686

Cover design by Alcorn Publication Design.
Index prepared by Diane Worden.
Printed in the United States of America.
Printed on recycled paper.

7
Measuring Government Performance

An Overview of Dysfunctional Responses

Pascal Courty
Gerald Marschke

Explicit performance measurement systems may elicit unintended and dysfunctional responses, also known as gaming responses. Understanding when such responses take place, their extent and their nature, is essential for improving the design of measurement systems and the overall effectiveness of performance incentives. This concern is reflected in the recent growth in empirical studies focusing on unintended behavioral responses to explicit incentives. We review this literature and try to provide a unifying framework to put into perspective the various classes of dysfunctional responses that have been identified in practice. We use this framework to discuss implications for the design of performance measurement systems.

The performance measure is the rule used to collect and aggregate the data generated by the agent's actions. The performance outcome is the value generated when that rule is applied to specific data. The next section proposes a formal classification of dysfunctional responses based on the terminology of the multitasking literature. Dysfunctional responses occur when the performance measure does not communicate correctly the marginal impact of decision making on the true objective of the organization. We distinguish three kinds of dysfunctional responses: 1) accounting manipulations, which are responses that boost the performance outcome but have no other impact on the organization; 2) gaming responses, which boost the performance outcome and have a negative impact on the organization; and 3) marginal misallocations, which have a positive impact on the organization but are suboptimal in the sense that alternative allocations would have a higher impact.

This classification is useful because it can help guide the organization's response to different dysfunctional behaviors. In the case of accounting manipulation, for example, the organization only has to invert the performance inflation relation and appropriately discount the rewards to performance achievement. If this cannot be done satisfactorily, however, accounting manipulations will have an indirect negative impact on the organization because the information contained in performance outcomes may be misinterpreted. Gaming responses should unambiguously be eliminated as they have both a direct negative impact on the organization (misallocation of resources) and an indirect one (misinterpretation of outcomes). Marginal misallocations often originate from the fact that the performance measure is too coarse and does not capture some dimensions of value added. The typical remedy is to complement the performance measure with finer measures or with alternative evaluation methods (i.e., subjective performance evaluation).[1]

After that we summarize the empirical literature on dysfunctional responses to performance measurement systems in public and private sector organizations, with an emphasis on the former. We then review the evidence on dysfunctional responses in the JTPA organization. Our point of departure is the earlier discussion of the weaknesses of the JTPA incentive system presented in Chapter 4. We exploit the analytical framework introduced there to understand the sources and consequences of dysfunctional responses. We conclude that section with some thoughts on the implications of the JTPA experience for WIA and its new performance incentive system.

The chapter ends with an assessment of the extent to which dysfunctional responses may impede the performance of measurement systems and draws lessons for policymakers. An important lesson of this review is that much progress has been made in identifying dysfunctional responses. A growing literature has produced studies that go beyond anecdotal reports and impressionistic evidence and try to identify dysfunctional behavior and measure performance inflation. Still, we find that this literature typically focuses on a narrow set of responses. In addition, the evidence reviewed rarely addresses the fundamental efficiency question of measuring the welfare impact of the dysfunctional responses identified. These conclusions suggest that much work is still necessary to further our understanding of dysfunctional behavior.

Before proceeding, we should acknowledge that others have discussed the existence of problems with performance measurement in both private and public organizations. See in particular Prendergast (1999), Propper and Wilson (2003), and Smith (1995). Other chapters in this book also discuss problems with performance measurement in JTPA and WIA. The main contribution of this chapter is to focus exclusively on the issue of dysfunctional responses, leaving aside more general problems associated with performance measurement, and to try to provide a comprehensive overview of such responses. To achieve this goal, we provide a theoretical framework to develop a formal classification of dysfunctional responses. This classification is useful to understand the practical challenges of identifying dysfunctional responses, to evaluate the negative impact of such responses, and to formulate appropriate remedies. We hope that this formal framework will be helpful in understanding the difficulties organizations face to correctly measure and reward productivity, and ultimately that it will support the design of more effective models of performance measurement and incentive systems.

DEFINITIONS OF DYSFUNCTIONAL RESPONSES

A central assumption of the incentive literature is that performance measurement influences behavior, and most importantly, that it may be sometimes difficult to anticipate how it does so. Performance measures encourage the right kind of behavioral responses only if they successfully communicate the organization's true objectives. In an early discussion of the subject, Blau (1955) warns that if performance measures are not perfectly aligned with the organization's objective, they may generate, in addition to intended responses, what could be called unintended or dysfunctional responses.

A dysfunctional response is an action that increases the performance measure but is unsupported by the designer because it does not efficiently further the true objective of the organization (see, for example, Kerr [1975] and Jensen and Meckling [1992]). The multitasking framework captures the notion that the investment allocation that maximizes performance outcomes does not necessarily correspond

to the allocation that maximizes value added (Baker 1992; Holmstrom and Milgrom 1991).[2]

Although all dysfunctional responses share the general property that they were not intended by the incentive designer, there are types of dysfunctional responses that correspond to different ways in which the measurement technology may be imperfect. To provide a more precise classification of dysfunctional responses, we borrow the language of the multitasking literature. The starting point of this literature is to assume that the agent invests in tasks. One could think of a task as a project. In the context of JTPA, for example, a task could be a single enrollee or a group of enrollees. The agent has to allocate resources across tasks and the issue is how the performance measure guides, or misguides, the agent's resource allocation. Each task is characterized by its type α. The agent privately observes the task's type α and invests in effort, e. The performance outcome for task α is

$$M_\alpha(e,g) = m_\alpha e.$$

We assume without loss of generality that $m_\alpha \geq 0$. Our specification ignores additive performance measurement noise.[3] This assumption is not restrictive for the analysis, which focuses on defining dysfunctional responses.[4] The principal's objective or social value added on task α is

$$V_\alpha(e) = v_\alpha e.$$

Finally, we assume that investment in effort is costly:

$$C_\alpha(e,g) = (\tfrac{1}{2})c_\alpha e^2,$$

where $c_\alpha \geq 0$. The performance outcome is

$$M = \Sigma_\alpha M_\alpha(e).$$

The fundamental assumption of the multitasking literature is that the principal can observe only M. The performance measure, however, is an imperfect measure of the agent's effectiveness because it aggregates the outcome of multiple tasks. As a result, the performance

measure may not be aligned with the true objective of the principal. Stated formally, this will be the case when the marginal return of effort on the measure is not the same as the marginal return of effort on the principal's objective, $m_\alpha \neq v_\alpha$.

Several comments are in order. First, this setup focuses exclusively on problems that are associated with the inadequacy of the measure to convey the true objective. It omits problems relating to the principal's ability to select the right measure and implement it properly and to the agent's willingness or capacity to respond to performance measurement. Although we briefly discuss the limitations imposed by these assumptions next, see Smith (1995) and Kravchuk and Schack (1996) for a more complete discussion of the problems that emerge when these assumptions do not hold. Some of these additional concerns regarding the implementation of performance measurement are also discussed in Chapters 3 and 4.

Second, the multitasking framework assumes that the principal's objective, $V_\alpha(e)$, is well defined. In practice, performance measurement sometimes fails because the principal's objective is poorly defined or because the principal must strike a compromise between potentially conflicting goals. Our analysis does not address these problems. Consider next our assumption that the agent chooses the resource allocation that achieves the highest outcome on the performance measures. This assumption rules out the possibility that the agent has his/her own preferences over resource allocation that conflict with performance measurement. It also rules out the possibility that the agent is actually an organizational unit composed of multiple decision makers with conflicting objectives, as is frequently the case in the real world. In addition, we assume that the agent understands the technology of production of the performance measure, $M_\alpha(e)$. Finally, our model abstracts from issues related to the dynamics of performance measurement. Here we consider a static model, ignoring the possibility that the principal may change the performance measure and the potential dysfunctional responses associated with such a possibility (Courty and Marschke 2003a).

Keeping these qualifications in mind, our setup suggests a formal definition of dysfunctional responses. A dysfunctional response is an investment choice that is different from the investment choice that max-

imizes the organizational goal. Formally, an agent who maximizes the performance outcome invests

$$e_\alpha = m_\alpha / c_\alpha$$

in task α, while the investment that maximizes the organizational objective is

$$\overset{*}{e}_\alpha = v_\alpha / c_\alpha.$$

A dysfunctional response occurs when $e_\alpha \neq \overset{*}{e}_\alpha$. We distinguish three types of dysfunctional responses:

1) Marginal misallocation: actions that enter the principal's objective but are distorted in the performance measure. Formally, $m_\alpha \neq v_\alpha > 0$ and $c_\alpha > 0$. To illustrate, consider the case of performance measurement in schools (Jacob 2005; Hannaway 1992). In recent years some policy analysts and public officials have advocated setting up performance measures for local school districts, possibly backed by educational subsidies as incentives (e.g., No Child Left Behind Act of 2002). Such performance measures are based on scores from standardized tests of reading, writing, and arithmetic. These tests do not measure the results of teaching citizenship, conflict resolution, and interpersonal skills—skills that are an important aim of primary schools. Because the tests do not measure citizenship, for example, the theory predicts that teachers will invest less, or possibly neglect altogether, this skill. Instituting performance measurement can produce distortions by causing agents to spend little time on activities that are productive but not fully taken into account in the performance measure.
2) Accounting manipulation: actions that increase the performance measure but do not enter the principal's objective and do not enter the cost function. Formally, $m_\alpha > 0$ and $v_\alpha = c_\alpha = 0$. Accounting manipulations are activities that boost the performance measure and do not waste resources. Such responses are informally known as "cooking the books" or "window dressing." Accounting manipulation increases the agent's chances of earning the rewards associated with higher perfor-

mance outcomes. They may not create welfare loss since the organization could, in principle, neutralize this behavior by appropriately discounting the rewards to higher performance outcomes (Courty and Marschke 2004). If such adjustment in rewards is possible, this class of dysfunctional responses does not have direct inefficiency implications (since $e_\alpha = e_\alpha$ on all tasks that actually enter the organization's objective). Often, however, the principal may not be aware of such responses. When this is the case, the informative power of the performance measure decreases and the principal may overreward those agents who invest more in accounting manipulation. Such dysfunctional manipulation would have indirect negative efficiency impacts on the organization.

3) Gaming: actions that increase the performance outcome negatively enhance the principal's objective and/or positively increase the cost function. Formally, $m_\alpha > 0$, $v_\alpha \leq 0$, and $c_\alpha \geq 0$, with at least one of the last two inequalities strict. The distinction between accounting manipulation and gaming is that the latter imposes a cost to the organization because the agent ends up wasting resources to boost performance. For example, if the activities involved in "cooking the books" waste resources, then they fall within the category of gaming. Gaming implies not only some kind of accounting manipulation but also a costly misallocation of resources.

This classification is useful because the organization's optimal response depends on the type of dysfunctional behavior under consideration. As mentioned earlier, the organization does not care about accounting manipulation if it can invert the performance inflation relation and appropriately discount the rewards to performance achievement. If this cannot be done satisfactorily, then accounting manipulations will have an indirect negative impact on the organization. Gaming responses should unambiguously be eliminated as they have both a direct negative impact ($e_\alpha \neq e_\alpha$) and an indirect impact on the organization. Marginal misallocations often originate from the fact that the performance measure is too coarse and does not capture some dimensions of value added. The typical remedy is to complement the performance measure with additional measures or with alternative evaluation methods (i.e., subjective performance evaluation).

REVIEW OF EVIDENCE OF DYSFUNCTIONAL RESPONSES IN PUBLIC AND PRIVATE SECTORS

Casual reports of dysfunctional responses to performance incentives abound. Although sometimes insightful, such accounts only give a very impressionistic view of the actual extent and impact of such responses. It is typically not possible to establish on the basis of such reports the amount of performance inflation that actually goes on, or to determine the existence of welfare loss. To draw relevant lessons for the design of performance measurement systems, one must develop systematic methods to identify and measure distortions. We start by discussing why it is difficult to systematically measure dysfunctional responses in practice. Next, we review different methods that have been successful at producing hard empirical evidence on dysfunctional responses.

Challenges in Identifying Dysfunctional Responses

Several difficulties arise when one tries to assess the extent of dysfunctional behavior. To start, demonstrating the existence of dysfunctional responses involves estimating relationships that are typically hidden from the researcher. One needs to identify actions that are not perfectly aligned with the principal's objective. Using the notation of the model, one needs to show that $m_\alpha \neq v_\alpha$, and also possibly that $c_\alpha > 0$, depending on the type of dysfunctional responses considered. But the researcher does not typically observe the marginal impacts of decision making on the production and cost functions.

To illustrate, consider the case of marginal misallocation. Researchers who study agent responses to performance measures often find evidence of actions that raise performance outcomes, but then find it difficult to demonstrate that these actions are suboptimal (i.e., to show that these responses are not the ones that maximize the stated objective of the organization). The difficulty lies in establishing the counterfactual of what the agent's value added would have been absent the agent's actions. Consider the cream skimming literature in job training programs that has studied enrollment responses to performance measurement in the JTPA organization (Chapters 6 and 9). Critics of JTPA's performance incentive system feared that the measures used,

which focus on labor market success (e.g., employment status) at the end of training, would encourage managers to enroll only those participants likely to perform well on employment measures—the most "job ready"—irrespective of how much they might gain from the program (that is, increase their human capital). Some studies have found evidence that program managers prefer the job ready, but this alone is not evidence of dysfunctional responses. To demonstrate dysfunctional response, one must show that the job-ready applicants are also those who do not benefit the most from the program (see Chapters 5 and 8, and Heckman, Heinrich, and Smith [2002]). Using our notation, although it seems intuitive that m_α/c_α is high for the job ready, one needs to prove that v_α/c_α is low for this target population to establish the existence of marginal misallocation.

The literature has circumvented this challenge by focusing on incentive schemes where dysfunctional responses can be unambiguously identified from the specifics of the contract. A substantial fraction of the literature focuses on accounting manipulation responses where the agent uses its discretion over the timing and reporting of performance outcomes to meet performance thresholds (Asch 1990; Courty and Marschke 2004; Healy 1985; Jacob and Levitt 2003; Oettinger 2002; and Oyer 1998). The advantage of focusing on such timing and misreporting strategies is that the observed responses can only be consistent with the specifics of the contract. Using the notation introduced earlier, these actions are unambiguously dysfunctional since $m_\alpha > 0$ while $v_\alpha = 0$. The main shortcomings of this approach, however, are that it can be applied only to a narrow set of dysfunctional responses and requires detailed information on the contracts and behavior that is often hidden from the researcher. Another shortcoming of this approach is that it will work only to identify accounting manipulations, to the extent that $c_\alpha = 0$, and such responses are likely to have lower direct efficiency impact than the other two types of dysfunctional responses.

A final approach to identify dysfunctional responses focuses on changes in performance outcomes that follow the introduction of a new performance measure. We discuss this approach in more detail in the next section.

Evidence of Dysfunctional Responses

Propper and Wilson (2003) present some evidence of dysfunctional responses in their review of the empirical literature on the use and usefulness of performance measures in the public sector. In this section, we review some of this evidence, as well as new evidence, using the classification presented above.

Health sector

Dranove et al. (2003) study whether the introduction of report cards changes how health care providers select patients. Report cards provide information about the performance of hospitals. Skeptics argue that health report cards may encourage providers to game the system by cream skimming, that is, by avoiding sick and/or seeking healthy patients. Their evidence shows that report cards led to substantial selection by providers, with a decline in the illness severity of patients, a finding consistent with a cream skimming hypothesis. They conclude that the overall impact of the report card was to reduce welfare. This evidence is consistent with marginal misallocation, since performance measurement generates a reallocation of resources that reduces efficiency.

Goddard et al. (2000) present a general discussion of the difficulties in implementing performance measurement. They consider the impact of the "Performance Framework," an initiative by the UK National Health Service to increase the importance attached to formal performance indicators in the health sector. They present qualitative interview evidence consistent with the hypothesis that performance measurement may generate a wide range of unintended responses.

School and training program

Jacob and Levitt (2003) investigate teacher cheating, a behavioral response consistent with accounting manipulation. Some school districts allocate school budgets on the basis of schools' performance. A number of highly publicized incidents of teacher cheating have fueled the suspicion that teachers have responded by "teaching the test" and manipulating students' grade-to-grade promotions to boost scores. However, most of this evidence is anecdotal. Jacob and Levitt propose

an innovative way to measure the extent of teacher cheating that combines measures on unexpected test score fluctuations and suspicious patterns of answers for students. They show that the joint distribution of these two variables should demonstrate systematic patterns if some teachers cheat and others do not.

Jacob (2005) presents some evidence of marginal misallocation. He examines the impact of an accountability policy implemented in the Chicago Public Schools in 1996–1997. He finds that math and reading achievement increased sharply following the introduction of the accountability policy. He also finds that teachers responded strategically to the incentives along a variety of dimensions—by increasing special education placements, preemptively retaining students, and substituting away from low-stakes subjects like science and social studies.

Oettinger (2002) presents a study of academic performance evaluations and shows that undergraduate students respond to nonlinear incentives. Due to the threshold effects implied by a discrete grade system, students tend to cluster slightly above grade boundaries. Using our terminology, this evidence is consistent with marginal misallocation because students strategically change effort decisions to meet performance thresholds, a behavior that is unlikely to be efficient. In fact, Oettinger's evidence suggests that the performance incentive generates allocations of effort over the duration of the term that depend on the realized grade history. The efficient allocation of effort, however, should not depend on grade history.

Burgess et al. (2003) evaluate the impact of a pilot incentive scheme in Jobcentre Plus, a large UK public job training agency, and present some evidence of marginal misallocation. The incentive scheme they consider gives team bonuses for five different targets that measure with varying degrees of precision the bureaucrats' effectiveness at placing the unemployed into jobs. The authors hypothesize that in such a multitasking environment, where different tasks are measured with different degrees of precision, workers may choose to exert effort on the tasks for which their actions are more easily verifiable. Consistent with this hypothesis, they find an impact on job placement (quantity measure) but little impact on less precise measures (quality measures).

Private sector: managers and salespeople

Explicit performance measures are commonly used in private sector occupations such as firm executives (using both accounting and stock market based measures) and salespeople.[5] Evidence of accounting manipulations from these occupations abounds. In an early contribution to the accounting manipulation literature, Healy (1985) documents that managers who are compensated for meeting annual income thresholds use their discretion over the timing of income reporting to smooth their compensation across accounting years. Similarly, Oyer (1998) uses differences in the date of the end of the fiscal year for companies that are otherwise similar to show that there is more variability in firms' sales at the end of the fiscal years—when salespersons' bonuses are computed—than in the middle. Oyer's evidence should be interpreted as accounting manipulation, if salespeople only manipulate sales reports. Alternatively, if salespeople reallocate effort over the accounting year, then the evidence should be interpreted as marginal misallocation.[6]

Gibbs et al. (2009) present some evidence that incentive designers are aware of marginal misallocations and actually structure measurement systems to address such problems. They use data on incentive contracts for auto dealership managers to investigate whether incentive designers internalize multitasking concerns. They show that incentive designers select pools of incentive measures that complement each other, and set the relative weights on the different measures selected to address multitasking concerns. For example, the extent to which a measure distorts incentives (by discouraging cooperation or encouraging a short-term focus) reduces the weight it receives. In addition, firms use additional performance bonuses, based on subjective performance evaluation, to balance multitasking and manipulation incentives.

EVIDENCE FROM JTPA

This section reviews the evidence of dysfunctional responses in the JTPA organization. Here we draw from our own work but also refer to the work of Heckman and Smith presented in Chapter 6 and Heinrich in Chapter 8 of this monograph. The evidence presented in this section

builds upon the characteristics of the JTPA incentive system presented in Chapter 4 (see also Courty and Marschke [2003b]).

We present two kinds of evidence. The first summarizes the econometric evidence of behavioral responses to JTPA's attempts to evaluate organizational performance, as well as estimates of the costs that are incurred when performance measures lead to dysfunctional behavior. This evidence establishes a clear link between the specifics of the incentive policies faced by bureaucrats and their behavior. The second kind of evidence is based on survey data of self-reported behavior. This evidence reviews a wider set of hypotheses regarding the implications of the incentive system but because the evidence is based on self-reported behavior, the inference is more anecdotal in nature and sometimes subject to interpretation.

Timing Strategies

Using data from the National JTPA Study, we document how in the first decade of JTPA, agencies delayed terminating unemployed enrollees, even after their training concluded, to maximize the performance outcomes (Courty and Marschke 1997, 2004). This strategic termination behavior can be of two types. The first type takes advantage of the fact that training agencies do not need to report the employment status of the enrollees who have completed their training on the date training ends but have a 90-day window to do so. Because labor market outcomes vary over time naturally on their own, training agencies have an incentive to strategically choose the date they report enrollees' employment outcomes. At the end of an enrollee's training, training agencies face a decision: terminate the enrollee and report her labor market outcomes or postpone termination in hopes that the outcome improves. The optimal termination strategy leads the training agency to terminate enrollees who are employed within the 90-day period following training either on the last day of training or on the first day of employment, whichever occurs first, and all others on the 90th day following training end. Courty and Marschke (2004) report evidence consistent with this hypothesis. They find that this strategic reporting increases the overall employment rate at termination, which was the most important performance measure at the time, by 11.3 percentage points, from 47.0 percent to 58.3 percent. Stated differently, training

agencies in their study would produce an employment rate outcome 20 percent lower if they were required to graduate enrollees (and report their performance outcomes) on the date they actually finish training.

The second type of strategic termination behavior takes place at the end of the fiscal year. Consider a stylized two-program-year incentive system where the training agency receives an award if the yearly labor market–based performance outcome exceeds a fixed performance standard. The training agency does not know its final aggregate performance outcome until the end of the program year because the labor market outcomes depend on random factors, such as the state of the local economy, which are outside its control. Because of the graduation strategy described above, the training agency reaches the end of the year with an inventory of enrollees who have finished training within the previous 90 days but are unemployed. At the end of the first program year, the training agency chooses how many from this inventory to graduate in the present program year, with the remainder to be graduated in the following program year. Assume there are n such persons, of whom n_1 will be graduated in the first program year and $n_2 = n - n_1$ in the next one. The training agency chooses n_1 to maximize the present value of the sum of the two awards.

The optimal graduation strategy on the last day of the first program year depends on the difference between the performance outcome and the standard as the last day arrives. Let $N = N_e + N_u$ be the number of persons who were graduated during the year (excluding the year's last day), where N_e and N_u and are the numbers of such persons graduated employed and unemployed, respectively. Let $\overline{S}$ be the performance standard. Three cases can be distinguished (see Figure 7.1). In case 1, on the last day of the year, the cumulative performance outcome exceeds the standard by so much that the training agency can graduate all unemployed enrollees. This corresponds to the *HIGH* region in Figure 7.1. In case 1, because $N_e/(N+n) \geq \overline{S}$, $n_1 = n$. In case 2, which corresponds to the *MED* region in Figure 7.1, the cumulative performance outcome exceeds the standard, but not by much. In case 2, because graduating all unemployed enrollees would push the outcome below the standard, it pays the training agency to graduate persons from its inventory only until the performance outcome equals the performance stan-

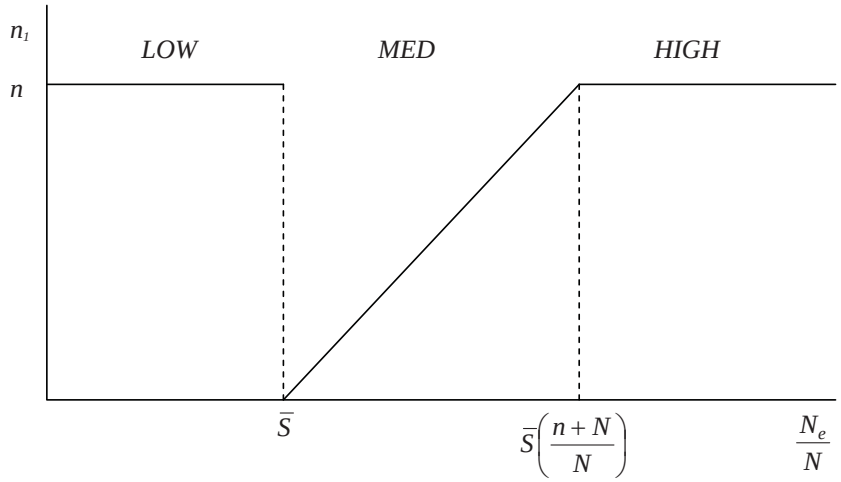
dard. That is, the training agency chooses n_1 such that $N_e/(N+n) = \bar{S}$. Rearranging yields

$$n_1 = \frac{N_e}{\bar{S}} - N.$$

This equation implies that n_1 lies between zero and n, approaching zero when the training agency just meets the standard and n when the training agency outperforms the standard by n/N percentage points or more. In case 3, corresponding to the *LOW* region in Figure 7.1, the training agency fails to meet the standard at the end of the year ($N_e/N \leq \bar{S}$). In this case, because it cannot win an award this year, the training agency "takes a bath," graduating all n persons from its inventory to maximize the probability of an award next year.

Courty and Marschke (2004) find evidence that training agencies pursued such a termination strategy. In particular, they found that JTPA training agencies delayed graduating idle, unemployed enrollees longer than idle, employed ones; graduated idle, unemployed enrollees sooner if they finished in the last three months of the program year than if they finished within the first nine months of the program year; and graduated

Figure 7.1 The Graduation Decision



unemployed enrollees who finished training in the last three months of the program year sooner if the training agencies were doing either very well or poorly relative to the employment standard. These findings are consistent with the two-period graduation model. Thus the evidence suggests that by timing performance measurement in this way, training agencies boosted their performance and their awards without providing higher-quality services or providing services more efficiently.

Courty and Marschke (2004) make the important distinction between responses that divert resources (e.g., agents' time) from productive activities and responses that simply reflect an accounting phenomenon. Using our terminology, the former would be labeled gaming and the latter accounting manipulations. Others have documented timing strategies but have not shown any efficiency impact. Courty and Marschke, however, provide evidence that the responses they identify, by consuming programmatic resources, have a negative impact on the true goal of the organization and thus conclude that these responses are more than a mere accounting phenomenon. They are evidence of gaming. For example, they find that earnings impacts are lower in those training agencies that engage more in termination strategies, which is consistent with the hypothesis that graduation timing is inefficient. In addition, they find that year-end timing is inefficient on two counts. First, training agencies are more likely to suddenly truncate training in June, an input distortion that is a direct consequence of the strategic manipulation of yearly performance that takes place in the month of June. Second, they find that earnings impacts are lower for those enrollees who receive training in June. They interpret this finding as evidence that training agencies substitute time and effort away from training toward the end of the program year.

Other Dysfunctional Responses

We review additional evidence also consistent with dysfunctional behavior. To start, we show that accounting manipulation behavior is not limited to the employment at termination performance measure by presenting evidence from other performance measures. Then we present evidence that the incentive system may also distort the enrollment decision and the training allocation decision.

Manipulating the wage and earning measures

The graduation decision was also influenced by the other class of JTPA performance measures: the performance measures based on wages and earnings. Courty and Marschke (2004) focus on the optimal termination strategy for employed enrollees. They show that training agencies may choose to terminate employed enrollees who have little chance of experiencing a wage increase, but wait on those employed enrollees who have a high likelihood of experiencing a wage increase. The training agency does not wait on all enrollees because by doing so it might lose credit for some employment. This risk is significant because approximately one-quarter of the enrolled who were employed on their graduation date were not employed at the same job three months later. The refined strategy that takes the wage measure into account implies that some employed enrollees should be terminated later than is predicted under the simple strategy presented earlier. Focusing on the enrollees who were employed at the end of training and who experienced a second employment spell under a new employer before the close of the 90-day window, Courty and Marschke show that those enrollees who experience a wage decrease are more likely to be terminated during the first employment spell. They also show that those participants who are graduated after the start of their secondary employment spell should experience higher wage offers. They conclude that the covariation of graduation delay and the wages in secondary employment spells appear consistent with a graduation strategy that maximizes the wage and earnings performance outcomes. This behavior qualifies as accounting manipulation or gaming depending on its efficiency impact.

Manipulating the follow-up measures

The switch from termination-based follow-up measures may have provoked several kinds of responses by local decision makers. First, the follow-up measures may have encouraged training centers to emphasize intensive training services (as opposed to employment-focused services, such as job search) in the hopes of producing larger and longer-lasting impacts on earnings and employment. Second, the follow-up measures may have induced caseworkers to extend their contact with enrollees beyond their termination dates. Courty and Marschke (2007) present some evidence suggesting that the follow-up performance measures

captured how effective training centers were at offering postprogram "quick fixes" (such as transportation allowances) to enable the enrollees to stay employed on the follow-up measurement date. These responses qualify as marginal misallocation because such quick fixes, although potentially productive, divert resources away from other activities that would have increased long-term earning and employment.

Manipulating the cost measure

In addition to employment and wage/earning measures, in the early years JTPA training centers faced a cost-based measure that judged the program's managers by how much they spent to produce an employment at termination. The incentives inherent in the cost and employment rate at termination measures were very similar (Courty and Marschke 2007). Thus, in time, JTPA officials came to believe that the cost measure also was encouraging short-run, "quick fix"–type job placement activities in lieu of longer-term activities with more training content. In 1990, when they replaced the termination-based measures with follow-up-based measures, they also phased out the cost measure because they believed "that the use of cost standards in the awarding of incentives [had] the *unintended effect* of constraining the provision of longer-term training programs" (italics ours) (*Federal Register* 1990). Such responses—if they indeed occurred—belong to the category of marginal misallocation.

Enrollment

Cream skimming—the use of the training center's considerable discretion to select enrollees on the basis of their expected effect on performance outcomes—is the core concern of most empirical analyses of JTPA's incentive system. The system judged training centers on the basis of postprogram employment and earnings levels, whereas the objective of training was skill development. Such performance outcomes induce training centers to choose persons with high levels of human capital at the expense of persons who would most benefit from training. The nature and evidence of cream skimming is discussed at length in Chapters 6 and 9.

Training selection

Researchers have also investigated the effects of JTPA performance incentives on training centers' training strategies. While training choice has received less attention than the cream skimming issue, its study is motivated by similar concerns—that short-term performance measures encourage training centers to emphasize "quick fixes," services that have no long-term impact on enrollee skills.

Marschke (2002) studies the effects of two performance measure reforms on the training strategies of JTPA training centers. In the early 1990s, the USDOL moved away from termination-based measures toward performance measured three months *after* termination. The USDOL also eliminated measures that rewarded training agencies that kept low the average cost of training an enrollee. Both reforms occurred in response to a growing perception that the training centers were relying heavily on job-placement-oriented services at the expense of more intensive kinds of training. Many policymakers also felt that the typical JTPA training spell was too short to be effective (average enrollment in the first decade of JTPA lasted only about five months).

Marschke (2002) finds that these performance reforms produced mixed results. The switch to performance measurement three months after training ends appeared to encourage agencies to offer the kinds of intensive training that raise the long-term earnings abilities of JTPA enrollees, but the impacts from this reform were offset by the elimination of the cost measure. Apparently the cost measure had been discouraging training agencies from offering classroom vocational training because it was the longest and most expensive of the major kinds of training. After the cost measure was removed, training agencies offered more classroom vocational training, but earnings impacts fell because classroom vocational training produced the smallest earnings impacts of the main kinds of training offered.[7]

In the context of the multitasking model, rewarding the employment rate at termination measure, for example, was leading the center to prescribe training activities that increased the training center's employment rate but reduced the earnings ability of JTPA enrollees. The employment rate at termination measure and earnings impacts are misaligned. The cost measure, on the other hand, was leading training centers to favor training types that increased both earnings impacts and the training cen-

ter's award. Thus, the cost measure, insofar as it discouraged the use of vocational training, a relatively low-gain and high-cost activity, appears to have been aligned. To conclude, this discussion presents some evidence consistent with the hypothesis that the choice of performance measure influences the agent's choice of resource allocation, and that different performance measures are imperfect in different ways in the sense that each produces different patterns of marginal misallocation.

General Test of Dysfunctional Responses

Courty and Marschke (2008) propose a different approach to identify dysfunctional responses. They look at how the correlation between the objective of the organization and the performance measure changes after a measure is introduced. The multitasking model predicts that this correlation should decrease after a measure is introduced if the agent engages in any kind of dysfunctional response. Courty and Marschke conclude that one can identify dysfunctional responses by estimating the change in correlation between a performance measure and the true goal of the organization before and after the measure has been activated. Using data from the JTPA incentive system, they test the hypothesis for the introduction of the follow-up measures, which corresponds to one of the most important changes in the history of JTPA's incentive system. They find some evidence consistent with the hypothesis and draw implications for the choice of performance measures.

WIA

WIA's performance incentive system is still relatively new and therefore little studied. Nonetheless, the record on JTPA may allow us to draw some conclusions about the performance of WIA's system. As Chapters 2 and 5 show, there are many similarities and some differences between performance incentives under JTPA and performance incentives under WIA. As in JTPA, most of the performance measures in WIA are based on enrollees' labor market outcomes. All labor market outcomes are measured after training ceases (as opposed to on the date of termination), as in latter-day JTPA. Moreover, WIA focuses on outcomes six months after job placement. In JTPA, performance outcomes were far more short term. This may be an improvement over

JTPA in two ways. First, because labor market outcomes are measured six months after an enrollee leaves the program, outcomes are harder to manipulate in WIA by the strategic timing of graduation. Second, measuring outcomes six months after graduation reduces the return to "quick fix"–type training strategies. As the JTPA evidence seems to show, longer-term measures lead to greater earnings gains from training. WIA also reinstates cost measures, which the evidence seems to show also improve earnings gains.

Interestingly, in its early years WIA includes among the JTPA-style performance measures a before-after measure of enrollees' earnings. Conceptually, the difference between an enrollee's earnings before enrollment and after termination is more similar to an earnings or employment gain—and thus more similar to the objective of job training under JTPA and WIA—than is a posttraining labor outcome. While it is more similar to a job training impact, a before-after earnings measure also suffers from potential problems. For the average enrollee, earnings dip just before entering job training, suggesting that her earnings would eventually rise even if the job training program had no value. This phenomenon, the so-called Ashenfelter dip, means that before-after earnings differences are distorted measures of the true impact of job training (see Heckman and Smith [1999] for a discussion of this point.) Thus, whether before-after measures lead to less cream skimming is a question that must be answered with empirical studies.

CONCLUSIONS

This chapter offers a comprehensive overview of dysfunctional responses covering theoretical concepts, empirical evidence from both the public and private organizations, and summarizing studies of dysfunctional responses that focus on the JTPA organization. We assess the extent to which dysfunctional responses may impede the performance of the measurement system, and we draw lessons for policymakers.

Taken together, the evidence suggests that performance measures elicit unanticipated responses because line workers and their managers gain a superior understanding of how to influence these measures. Managers and workers acquire through their day-to-day operation of their

programs an expert's knowledge of the levers available to manipulate performance outcomes. Because the designers of the performance measures are remote from the everyday operations of the agencies they oversee, they lack this knowledge. This information asymmetry means that the designers of performance measures cannot anticipate all behavioral responses *ex ante*. An implication is that dysfunctional responses can present substantial challenges to the design of performance measurement systems. At least three lessons can be drawn from the evidence.

1) Designers of performance measures should consider how local decision makers respond to the performance measures and accept that they cannot anticipate all responses. Most performance measures elicit unanticipated responses because agents gradually gain a superior understanding of how to influence them. Designers should encourage some of these unanticipated responses and discourage others.
2) Designers should respond differently to different types of dysfunctional behavior. In the case of accounting manipulation, for example, the organization only has to appropriately discount the rewards to performance achievement. Gaming responses should unambiguously be eliminated as they have both a direct negative impact on the organization (misallocation of resources) and an indirect one (misinterpretation of outcomes). Marginal misallocations often originate from the fact that the performance measure is too coarse and does not capture some dimensions of value added. The typical remedy is to complement the performance measure with finer measures.
3) The evidence reinforces the conjecture that explicit performance measures impose costs—monitoring and improvement consume resources (Prendergast 1999). Until performance measure designers discover them, the dysfunctional responses that imperfectly conceived performance measures engender can undermine the organization's mission. These costs make the use of performance measurement systems uneconomical for many public sector organizations because they raise more management problems than they solve. This may explain why such organizations rarely implement explicit performance measures.

In a separate line of research, we characterize a process by which the designers of performance measures learn about and respond to local decision maker responses (Courty and Marschke [2003a]; see also Heinrich and Marschke [2010]). This feedback loop suggests that performance measurement systems must be continuously monitored and improved. We conclude that implementation is not a static, one-time challenge but a dynamic one.

An important lesson of this review is that much progress has been made in identifying dysfunctional responses. A growing literature has produced studies that go beyond casual reports and anecdotal evidence and try to measure performance inflation and assess the impact of these responses on the organization. Still, we find that this literature typically focuses on a narrow set of responses. In addition, the evidence reviewed rarely compellingly addresses the fundamental efficiency question of measuring the welfare impact of the dysfunctional responses identified. This suggests that there is still much work to be done to further our understanding of dysfunctional behavior.

Notes

We would like to thank James Heckman and Carolyn Heinrich.

1. Subjective performance evaluation is based on judgment and is more qualitative and flexible than explicit performance incentives. Unlike explicit performance measures, subjective performance measures cannot be verified by outside parties and therefore organizations cannot contract upon them.
2. The theoretical literature on incentive provision is reviewed in Gibbons (1997) and Prendergast (1999). Marschke (2001), Courty and Marschke (2003a), and Propper and Wilson (2003) also review this literature and draw implications specific to government organizations. Dixit (2002) and Burgess and Ratto (2003) review the broader literature on organizational design and focus on issues that are specific to the public sector. Interestingly, this framework was introduced to explain why high-powered explicit incentives, whose canonical illustration is a piece rate system, appeared far less frequently in practice than is predicted by standard principal-agent models. The point of the theoretical literature on multitasking was to extend the principal-agent model to accommodate the possibility that high-powered explicit incentives may not be optimal when the principal cannot perfectly measure the objective she wants the agent to pursue.
3. Our specification is very similar to the specification in Baker (2002), who assumes that $V = f.a + \varepsilon$ and $P = g.a + \phi$, where f and g are vectors of marginal products

of actions, a, in the principal's objective and performance outcome equations. We ignore the error terms ε and ϕ as they do not influence the agent's action choice (see the following note).

4. In the standard principal-agent model, measurement noise plays a role in the determination of the optimal contract, but it does not directly influence the agent's investment decisions.
5. When output can be measured, manual work is another occupation where explicit performance measurement is common. In an unusual study, Lazear (2000), for example, discusses how an installer of automobile glass minimized the impact of potential dysfunctional responses to the introduction of piece rate rewards.
6. Asch (1990) also presents some evidence of reporting timing. She shows that navy recruiters who receive awards for meeting year-end recruitment quotas respond by reallocating their work efforts over the year.
7. This finding is consistent with the results of the National JTPA Study, which found that compared with job search assistance and on-the-job training, vocational classroom training produced the weakest earnings and employment gains (see Orr et al. 1994). One interpretation of Marschke's finding is that the deactivation of the cost measure in the early 1990s was misguided. The finding does not rule out, however, that at the same time it discouraged the use of vocational classroom training, the cost measure encouraged training centers to cut corners in the delivery of services, or to dilute or prematurely shorten training activities for enrollees.

References

Asch, B. J. 1990. "Do Incentives Matter? The Case of Navy Recruiters." *Industrial and Labor Relations Review* 43(3): S89–S106.

Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100(3): 598–614.

———. 2002. "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources* 37(4): 728–751.

Blau, P. 1955. *The Dynamics of Bureaucracy: A Study of Interpersonal Relations in Two Governmental Agencies*. Chicago: University of Chicago Press.

Burgess, Simon, Carol Propper, Marisa Ratto, and Emma Tominey. 2003. "Incentives in the Public Sector: Evidence from a Government Agency." CMPO Working Paper No. 03/080. Bristol, UK: Centre for Market and Public Organisation, University of Bristol.

Burgess, Simon, and Marisa Ratto. 2003. "The Role of Incentives in the Public Sector: Issues and Evidence." *Oxford Review of Economic Policy* 19(2): 250–267.

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

———. 2003a. "Dynamics of Performance Measurement Systems." *Oxford Review of Economic Policy* 19(2): 268–284.

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

———. 2004. "An Empirical Investigation of Dysfunctional Responses to Explicit Performance Incentives." *Journal of Labor Economics* 22(1): 23–56.

———. 2007. "Making Government Accountable: Lessons from a Federal Job Training Program." *Public Administration Review* 67(5): 904–916.

———. 2008. "A General Test for Distortions in Performance Measures." *Review of Economics and Statistics* 90(3): 428–441.

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

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite. 2003. "Is More Information Better? The Effects of 'Report Cards' on Health Care Providers." *Journal of Political Economy* 111(3): 555–588.

Federal Register. 1990. 55(4). Washington, DC: U.S. Government Printing Office.

Gibbons, R. 1997. "Incentives and Careers in Organizations." In *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress*, David Kreps and Kenneth Frank Wallis, eds. New York: Cambridge University Press.

Gibbs, Michael, Kenneth Merchant, Wim Van der Stede, and Mark Vargus. 2009. "Performance Measure Properties and Incentive System Design." *Industrial Relations*: *A Journal of Economy and Society* 48(2): 237–264.

Goddard, Maria, Russell Mannion, and Peter C. Smith. 2000. "Enhancing Performance in Health Care: A Theoretical Perspective on Agency and the Role of Information." *Health Economics* 9(2): 95–107.

Hannaway, Jane. 1992. "Higher Order Skills, Job Design, and Incentives: An Analysis and Proposal." *American Educational Research Journal* 29(1): 3–21.

Healy, P. 1985. "The Effect of Bonus Schemes on Accounting Decisions." *Journal of Accounting and Economics* 7(1–3): 85–107.

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

Heckman, James J., and Jeffrey Smith. 1999. "The Pre-Program Earnings Dip and the Determinents of Participation in a Social Program: Implications for Simple Evaluation Strategies." *Economic Journal* 109(457): 313–348.

Heinrich, Carolyn, and Gerald Marschke. 2010. "Incentives and Their Dynamics in Public Sector Performance Management Systems." *Journal of Policy Analysis and Management* 29(1): 183–208.

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

Jacob, Brian A. 2005. "Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89(5–6): 761–796.

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

Jensen, M. C., and W. H. Meckling. 1992. "Specific and General Knowledge and Organizational Structure." In *Contract Economics*, L. Werin and H. Wijkander, eds. Oxford, UK: Blackwell, pp. 252–271.

Kerr, Steven. 1975. "On the Folly of Rewarding for A While Hoping for B." *Academy of Management Executive* 18(4): 769–783.

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

Lazear, Edward. 2000. "Performance Pay and Productivity." *American Economic Review* 90(5): 1346–1361.

Marschke, Gerald. 2001. "The Economics of Performance Incentives in Government with Evidence from a Federal Job Training Program." In *Quicker, Better, Cheaper? Managing Performance in American Government*, Dall W. Forsythe, ed. Albany, NY: Rockefeller Institute Press, pp. 61–97.

———. 2002. "Performance Incentives and Organizational Behavior: Evidence from a Federal Bureaucracy." Working paper. Albany, NY: University at Albany, State University of New York.

Oettinger, Gerald S. 2002. "The Effect of Nonlinear Incentives in Performance: Evidence from 'Econ 101.'" *Review of Economics and Statistics* 84(3): 509–517.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Winston Lin, George Cave, and Fred Doolittle. March 1994. *The National JTPA Study: Impacts, Benefits, and Costs of Title II-A*. Bethesda, MD: Abt Associates Inc.

Oyer, P. 1998. "Fiscal Year Ends and Non-Linear Incentive Contracts: The Effect on Business Seasonality." *Quarterly Journal of Economics* 113(1): 149–185.

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

Propper, Carol, and Deborah Wilson. 2003. "The Use and Usefulness of Performance Measures in the Public Sector." *Oxford Review of Economic Policy* 19(2): 250–267.

Smith, Peter. 1995. "On the Unintended Consequences of Publishing Performance Data in the Public Sector." *International Journal of Public Administration* 18(2–3): 277–310.