THE PERFORMANCE PARADOX IN THE PUBLIC SECTOR

SANDRA VAN THIEL
Erasmus University Rotterdam, the Netherlands
FRANS L. LEEUW
Dutch Educational Review Office

Nowadays, states spend more attention, time, and money on performance measurement and evaluation in the public sector than ever before (Organization for Economic Cooperation and Development [OECD], 1996; Pollitt & Bouckaert, 2000, p. 87; Power, 1997). Results-based management is the talk of the day at all levels of the public sector: local, regional, national, and even supra national. Schools and universities, local governments, and other administrative agencies, also developmental aid organizations (nongovernmental organizations and international nongovernmental organizations) and organizations such as the World Bank, are all involved in producing data and information on performance results and—if possible–impact. Power (1994, 1997, 2000) even refers to the "audit explosion" or the "audit society." Believers in New Public Management (NPM) attribute a high priority to measuring output and outcomes and aim to base their new policies and management activities on this type of information—ideally meant to make policy implementation more efficient and effective. However, evaluation studies show that many attempts to introduce results-based management are still unsuccessful (see, for example, Leeuw & Van Gils, 1999, for a review of Dutch studies). Nevertheless, the need for measuring output, outcomes, and evaluation activities remains an important element in statements by politicians and administrators focused on improving government's performance.

Below, we will argue that this increase of output measurement in the public sector can lead to several unintended consequences that may not only invalidate conclusions on public sector performance but can also negatively influence that performance. We will show that a number of characteristics of the public sector can be counterproductive to developing and using performance indicators, illustrated by different examples. Finally, we will conclude with some suggestions on how to deal with the problem of performance assessment in the public sector. We believe this question is important

Public Performance & Management Review, Vol. 25 No. 3, March 2002 267-281

because—although not without problems—performance measurement indeed can be of value to the public sector.

Performance Assessment in the Public Sector

The increased attention to performance assessment in the public sector coincides with the rise of administrative reform (cf. Power, 2000). In the 1980s, economic decline and increased international competition triggered such reform in most western states. *New Public Management* was the catchword (Hood, 1994). The objective was twofold: to cut budgets and to improve the efficiency and effectiveness of government bureaucracy. To achieve the latter objective, market-type mechanisms such as privatization, competitive tendering, and vouchers were introduced in the public sector, and departmental units were hived off into quasi-autonomous nongovernmental organizations (quangos). Examples can be found everywhere (for a review of 10 OECD countries, see Pollitt & Bouckaert, 2000).

The practitioner theory underlying these changes is that politicians should stick to their core business, that is, developing new policies to realize (political) goals. Osborne and Gaebler's (1992) adage was "steering not rowing." According to these NPM gurus, policy implementation should be left to the market or, if that is not possible, to (semi)-autonomous organizations operating in a quasi-market environment (e.g., competition between schools or hospitals). This separation of policy and administration is facilitated through contracts being drawn up between the government and the organization that implements the policy. The contracts articulate which task has to be carried out and what the executive agent will receive as a "reward." The agent's performance is expressed in terms of performance indicators, such as the number of goods or services rendered. Input management is thus replaced by a results-based orientation. Similar changes took place within government bureaucracy as well, where self-management and contract management were introduced to (partly) replace hierarchical steering.

The aforementioned changes in the public sector led to the adoption of a large number of private sector techniques to measure and improve performance, such as performance indicators. Not only do indicators enable politicians to measure and evaluate the performance of public and private policy-implementing organizations, they also increase the opportunities to account for performance—another important goal of administrative reform (Jenkins, Leeuw, & Van Thiel, in press). Obviously, all these changes were fed by a strong belief in the measurability of performance in the public sector. However, as we shall argue below, that belief may have been somewhat simplistic (cf. Fountain, 2001).

Unintended Consequences

Approximately 20 years after the first reforms were initiated, it has become clear that performance assessment in the public sector is not without problems or unintended consequences (Leeuw, 2000; OECD, 1996). The audit explosion has led to a strong increase in the number of regulators and auditors (cf. Leeuw, Toulemonde, & Brouwers's [1999] survey among evaluators in Europe).[1] Within central government,

the number of evaluative organizational units or divisions has increased as well (cf. Algemene Rekenkamer, 2000, on the Dutch case). The expenditure on these audit-type activities has increased as the attention to the evaluation capacity or the evaluation infrastructure of public sector organizations has grown.

Although all these developments undoubtedly focus on the production of relevant and empirically correct information about the performance and accountability of these organizations, there is more and more evidence that they also have unintended and even undesirable side effects. Or, as Schwarz puts (in press) it,

> There is a desire to supply managers, policymakers, legislators and the general public with evaluative information that is perceived to be reliable, valid and credible. Evaluative information that lacks these characteristics stands little chance of enhancing transparency, accountability and democratic governance. Yet, mechanisms for assessing the "quality" (i.e. reliability, validity, credibility) of evaluative information conjure up perverse images of what has been termed an audit society characterized by increasing layers of inspection, audit, evaluation and assessment. The audit society expends a huge amount of resources in assurance activities whose most immediate consequence is to increase bureaucratization.

There are different reports of examples of these unintended consequences. In 1991, Bouckaert and Balk wrote about 13 diseases of public productivity measurement. These diseases were the result of wrong assumptions underlying measurement, measurement errors, and problems concerning the content, position, and amount of measures. The authors wondered whether it is indeed possible, desirable, or even necessary to measure public sector performance (Pangloss disease) because "government is efficient, because if it is not efficient, why hasn't it already been changed?" (Bouckaert & Balk, 1991). People can get disoriented about public sector performance as a result of measurement. For example, "Northern Great Britain seems to have more fires than other European countries because it has a better statistical technique for measuring." Bouckaert and Balk call this the Mandelbrot disease. They call for a management of the meaning of measurement.

Similarly, Smith (1995) wrote about eight unintended consequences of monitoring and investigating (auditing) performance. For example, the use of performance indicators can inhibit innovation and lead to ossification, that is, organizational paralysis. Another effect is referred to as tunnel vision, which "can be defined as an emphasis on phenomena that are quantified in the performance measurement scheme at the expense of unquantified aspects of performance" (p. 284). Other unintended side effects are suboptimization, which is defined as "narrow local objectives by managers, at the expense of the objectives of the organization as a whole" (p. 287) and measure fixation, "an emphasis on [single] measures of success rather than [on] the underlying objective" (p. 290). Fitz-Gibbon (1997, pp. 87-95) used Smith's approach in a survey among 104 head teachers of primary schools in the United Kingdom. She found that "with the exception of ossification, each of these possibilities [i.e., unintended side effects] was commented upon by head teachers in open-ended items on the questionnaires. These are not theoretical problems but *actual* [italics added], already-perceived problems" (p. 87).

Other examples are found easily. In the education sector, for example, Rand researchers Brewer, Gates, and Goldman (cited in the January 21, 2001, edition of the *Times* Higher Education Supplement **[PLS PROVIDE REF]**) argue that in the United States, too much focus on criteria, benchmarks, and other evaluative tools has led to mediocre institutions mimicking the outward appearance of prestigious universities rather than improving their teaching. A recent British evaluation of the higher education quality assessment and assurance activities in the United Kingdom refers to behavioral costs involved in producing performance information (Higher Education Funding Council for England, 2000; cf. Ghoshal & Moran, 1996). The British Quality Assurance Agency for higher education is currently under attack by universities and lecturers, primarily because the criteria used for quality and performance are believed to be inflexible, focusing too much on administrative topics and sometimes just simply wrong.

Bayer (personal communication, 2000)[2] mentions that higher education institutes that are unjustly given a high rank (because of measurement errors) will start to attract more highly qualified students and in the longer run indeed improve their performance. However, the opposite might also be true. Light (1993, p. 223) gives an example in a study on inspectors general (IGs) within the U.S. federal government. He found that "as the IG numbers go up, effectiveness may go down" (p. 223).

Other instruments of performance assessment are also debated. For example, Schmidtlein (1999, p. 1) concludes that very few studies have been carried out on the effectiveness of performance budgeting although most state authorities in the United States are using it (37 states in 1998). Although government officials claim to manage schools based on their output, school boards and directors claim otherwise and continue as before. Schmidtlein compares this situation to a black market and concludes that "any system employing financial incentives that are expected to achieve selective outcomes is highly experimental and will almost certainly create unintended consequences, some of them negative" (p. 11).

Finally, Goodlad (2000, pp. 71-79) points to another problem of performance assessment. Whereas "huge quantities of people's time and effort have been devoted to quality assurance mechanisms in education, often [there is] no reference to what exactly is . . . assured. The quality assurance documentation has not been very illuminating about what quality actually was." He refers to the "dangers of templates and benchmarking such as that of 'dumbing down' by defining subjects in terms of minimal achievements with the associated danger of teaching to the syllabus."

Reviewing the aforementioned and other studies on this topic over the past 5 to 10 years (cf. Leeuw, 2000), we conclude that it is reasonable to assume that no matter how well intended evaluation and auditing are, they have also led to some unintended consequences. First of all, the proliferation of regulators and auditors has increased the monitoring costs of the organizations and the state. Second, within policy-implementing organizations, the increased measure pressure can create dysfunctional effects such as ossification, a lack of innovation, tunnel vision, and suboptimization. These unintended effects can jeopardize the effectiveness and efficiency of policy implementation. Third, there is some evidence that monitoring has led to symbolic behavior; that is, monitoring appears to be in place but is in fact not. And finally, in some cases it is unclear what is actually being measured (e.g., the definition of quality).

These findings raise the questions of how to detect these unintended effects, how to prevent them, and how to explain them. Explanatory social and behavioral theories can be of help here. These theories can also help us to understand why some institutional conditions lead to more *effects pervers* than others. To that end, we will discuss one of the unintended consequences more in depth: the performance paradox.

The Performance Paradox

The performance paradox refers to a weak correlation between performance indicators and performance itself (Meyer & Gupta, 1994; Meyer & O'Shaughnessy, 1993). This phenomenon is caused by the tendency of performance indicators to run down over time. They lose their value as measurements of performance and can no longer discriminate between good and bad performers. As a result, the relationship between actual and reported performance declines.

Deterioration of performance indicators is caused by four processes (Meyer & Gupta, 1994, pp. 330-342). The first process is called positive learning; that is, as performance improves, indicators lose their sensitivity in detecting bad performance. In fact, everybody has become so good at what they do that the indicator becomes obsolete. The second process is called perverse learning. When organizations or individuals have learned which aspects of performance are measured (and which are not), they can use that information to manipulate their assessments. For example, by primarily putting all the efforts into what is measured, performance will go up. However, overall there may be no actual improvement or perhaps even a deterioration of (other aspects of) performance (cf. tunnel vision) (Smith, 1995). The third process, selection, refers to the replacement of poor performers with better performers, which reduces differences in performance. Only good performers remain, and the indicator loses its discriminating value—almost resembling a consequence of the survival of the fittest mechanism. And fourth, suppression occurs when differences in performance are ignored (see below for an example).

It is important to understand that the paradox is not about performance itself but about the reports on performance. Contrary to the expectation, indicators do not give an accurate report of performance. This could mean that performance is worse than reported (overrepresentation) but also that it is better than reported (underrepresentation). In the latter case, the performance paradox might be considered harmless. However, when the results of performance assessment are used to evaluate organizations or persons, situations can arise where these are unjustly sanctioned. An example given by Wiebrens and Essers (1999) can illustrate this. These authors report on the percentage of crimes solved by the Dutch police.

The percentage of crimes solved is decreasing, indicating that the police's performance is deteriorating. However, during the time period studied, more perpetrators have been arrested, prosecuted, and penalized than before, which would indicate an improvement of performance. Wiebrens and Essers (1999) show that crime patterns in the Netherlands have developed in a way that invalidates the (internationally well-established) indicator. For one, crime has become more violent, but the indicator does not differentiate between, for example, felonies and misdemeanors. Moreover, more groups of criminals have been arrested committing a crime together such as vandalism,

which reduces the average number of crimes per criminal. Wiebrens and Essers conclude that it is not the police that are performing badly but the indicator and that it therefore should be replaced.

An example of a performance paradox in a case of overrepresentation is taken from Smith (1995). In the British National Health Service, it was agreed that patients should be on a waiting list for an operation no longer than 2 years. This measure appeared successful, as the average waiting time decreased. However, on further inspection it was found that—the waiting time only beginning after the first hospital consultation—consultation was postponed to decrease the waiting time (perverse learning). In fact, the average waiting time did not decrease at all but was merely shifted in time. The indicator did not accurately reflect performance; it reported an improvement where there was none.

The two examples show that a performance paradox can be evoked unintentionally (the police example) or deliberately (the health care example). The causes for these two types of paradoxes are different.

Unintended Performance Paradox

An unintended performance paradox can, for example, be the result of minimal accountability requirements. Studies into the accountability of quangos[3] in the United Kingdom (Hall & Weir, 1996) and the Netherlands (Van Thiel, 2001) show that few requirements are imposed, and not in all cases. And if accounts of performance are given by public sector organizations, targets and benchmarks are often missing. Analysis of annual reports shows that output and input indicators are used most often, but productivity indicators, quality measurements, and cost prices are still notoriously absent (Van Thiel, 2001).[4]

The fewer the number of performance indicators, the more difficult it becomes to obtain an accurate report of the performance (Meyer & Gupta, 1994). Moreover, evaluation of auditors' reports shows a strong inclination to focus on procedures rather than actual performance and on the absence or presence of performance indicators rather than their quality or content (Leeuw, 2000; OECD, 1996).

A second cause of an unintended performance paradox is the elusiveness of policy objectives (Wilson, 1989, pp. 32-33). Public policies often have many, and sometimes contradictory, goals. Consequently, performance indicators are usually not neutral but contested measures in the public sector, both between politicians and between politicians and managers (McGuire, 2001). This ambiguity complicates the evaluation of the efficiency and effectiveness of policy implementation, for it is difficult to determine which objectives are most important and to whom. Performance indicators can thus—unwittingly—cause dysfunctional effects. An example on labor exchange agencies will illustrate this.

On one hand, labor exchange agencies are charged to help those clients who are most in need of their services, such as uneducated or poor people (outcome). On the other hand, their performance is based on the number of successful "transactions," that is, helping somebody to find a job again (output). The success rate will be highest when taking on only clients who have good chances—most likely not the uneducated

and poor ones. The performance standards thus do not reinforce or measure the policy objectives. In a study on American local job-training agencies, Heinrich (1999, pp. 369-371) found that local program administrators developed margins or buffers in their performance standards to reduce the risk of failing to meet performance requirements. Bouckaert and Ulens (1998) report that some Belgian welfare services have adapted their policy; the "difficult" clients are skipped in favor of the "easy" ones (see also below on cream skimming or cherry picking). This behavior is the unintended result of the indicators in use (cf. also Heckman, Heinrich, & Smith, 1997, p. 393).

Thirdly, policy goals are often nonquantifiable and hard to measure. For example, how do we measure national safety or national health? Does an increase in the number of apprehended criminals make us feel safer or less safe (cf. Bouckaert & Balk, 1991)? And will an increase in the number of medical operations in hospitals make us feel more healthy or less? Mol (1996) shows that most performance indicators in use by the National Logistic Command of the Dutch Armed Forces do not refer to primary processes but only to internal management affairs—probably because those are easier to develop. Meticulously prepared reports are therefore put to limited use. Mol explains this observation from the traditional management orientation in the army:

> Management control . . . in fact remains focused—to some extent at least—on hierarchical subordination of commanding officers to their superiors. This orientation is reflected in the application of indicators which are irrelevant and incoherent from the point of view of performance measurement. Targets may be aimed at the enforcement of some compliance to central authority or regulations, and thereby focus on the way in which the units are managed internally . . . rather than performance controls. Management reports will thus focus on . . . deviances directly rather than on their explanation in relation to any criterion of economy, efficiency or effectiveness. (pp. 80-81)

Measurement problems make it difficult to be certain about the validity of performance assessment. Because many policy-implementing organizations are monopolists, there is no comparative information or benchmarks to make evaluation of performance possible as well.

Finally, the chance of an unintended performance paradox occurring is enhanced by a strong emphasis on monitoring and efficiency within the organization. Extensive use of performance indicators can create a situation in which agents learn what aspects of their work are important to the principal. Smith (1995) labels this measure "fixation." By increasing the efficiency and effectiveness of those aspects, agents are assured of the principal's approval (gaming) (Smith, 1995). Such learning effects and anticipatory behavior of agents could, in the end, lead to ossification. Executive agents will aim all their efforts at meeting the principal's monitoring demands. It is then no longer relevant whether policy implementation is taking place in the most efficient manner, just as long as it appears to be efficient (cf. suppression and symbolic mimicking above).

The following example illustrates the effect of measure fixation. A survey by *USA Today* (2000) and the American Federation of Teachers found that the introduction of standards on what students need to learn led to teachers helping their pupils to cheat. Schools have become obsessed with the tests. Teachers outline "exactly what they

want students to learn, preparing curricula that concentrate mainly on what's expected to be tested. **[PLS PROVIDE PAGE]**" Of course, most educators will not call it cheating but simply giving extra help to a failing student. Why do teachers do this? The answer is very simple: Student scores determine school budgets, teacher salaries, and principals' job positions. The better the scores, the more budget a school has. A school that does not prepare its students for the tests runs the risk of losing budget. One can doubt, however, whether this cheating helps students' intellectual development and training, which was the aim of introducing the standards to begin with. It may perhaps even increase the risk of dropouts.[5] An increased attention for and use of performance indicators in the public sector—as stimulated by NPM—may thus unintentionally increase the risk of the occurrence of a performance paradox.

Deliberate Performance Paradox

Public sector organizations may also decide to evoke a performance paradox on purpose. For example, they can "sabotage" an audit when they consider it an act of distrust (Ghoshal & Moran, 1996; Leeuw, 2000). Delay or other noncooperative behavior does not necessarily impair efficient and effective policy implementation but contaminates the relation between reported and actual performance.[6]

The agent can try to hide ill performance by misrepresenting or misinterpreting performance indicators (cf. perverse learning).[7] A number of the aforementioned unintended consequences as listed by Smith (1995, see above) fit into this strategy. For example, agents can emphasize easily quantifiable aspects of performance (tunnel vision) and leave out reports on aspects of policy implementation that are difficult to measure. Also, they can confine themselves to reporting on the performance of parts of the organization, preferably those parts that are most efficient (suboptimization), or on short-term objectives (myopia). In this context, LeGrand and Bartlett (1993, pp. 31-34) refer to cream skimming. Cream skimming, or cherry picking, is the tendency of executive agents to discriminate against inefficient aspects of the policies to be implemented by providing services or goods only to those who make the least or least expensive use of them (e.g., in health care, chronically ill patients are excluded). Cream skimming makes the organizations appear to be more successful than they actually are. The example on the labor exchange organizations above illustrates the effects of cream skimming.

Even performance assessment itself can suffer from the aforementioned problems. For example, evaluation studies of organizations or policy programs can suffer from partiality due to the selection of data sources, respondents, and research methods. Also, it is not uncommon that evaluation studies are not translated into policy or management changes. The meta-evaluation of developmental aid program evaluations by the World Bank (1998) gives some nice examples in this respect.

In sum, unintended and deliberate performance paradoxes occur for different reasons and are the results of different circumstances. However, a deliberate paradox can occur only if the conditions for an unintended paradox are present as well. This is evident in the labor exchange agency; cream skimming can occur because the performance indicators do not reinforce the policy objectives.

The Occurrence of a Performance Paradox in the Public Sector

An argument could be made that certain characteristics of the public sector increase the chance of a performance paradox's occurring (cf. Fountain, 2001). First, a performance paradox is the result of a discrepancy between the policy objectives set by politicians and the goals of executive agents (Smith, 1995). The translation by managers of ambiguous, nontangible policy objectives into operational goals leaves room for deviations in policy implementation, which can lead to a performance paradox. It should be noted, though, that in some cases such discretionary authority is given intentionally, either because politicians want to appease multiple stakeholders or to facilitate executive agents' work (cf. Torenvlied, 2000). This illustrates that policy objectives are not neutral but contested both among politicians as well as between politicians and managers (McGuire, 2001).

Other factors that could increase the chance of a performance paradox in the public sector are the lack of potential bankruptcy (dismissal) of public sector organizations and the disjunction between costs and revenues (LeGrand, 1991). Many public sector organizations have problems in estimating the exact costs of their products and services. That is why in the Netherlands, public sector organizations are, as yet, not allowed to undertake commercial activities. The risk of cross-subsidizing is too high; that is, the development of new, commercial activities profits from the experience and knowledge achieved through the publicly paid activities (Commissie Cohen, 1997). One of the causes of the problem of calculating the costs of products and goods is that in the public sector, production and consumption of goods often occur at the same time:

> Public servants encounter constituencies whose preferences are ambiguous, dynamic and shaped significantly by and through their relationship with the public bureaucracy itself. Political bureaucrats have an obligation to do more than satisfy customers. They must identify and aggregate preferences in ways that sustain political legitimacy and minimize political inequality. (Fountain, 2001, p. 65)

All these factors create leeway for public sector organizations to manipulate information and get away with it because there are no real sanctions. Policy-implementing agencies have the advantage of their expert knowledge, both about the policy and its implementation. Without such information, government will be unable to prove manipulation or misrepresentation of performance information. Finally, many executive agents in the public sector are monopolists.[8] Hence, there is no comparative information to evaluate the performance, but there are also no immediate substitute agents to which the government can turn if it is dissatisfied with the agent's performance—except at a very high cost.[9]

Detection and Prevention of a Performance Paradox

Although most readers will have recognized the examples we have given so far, it is not easy to trace a performance paradox in progress. Not only can it take on many different forms, it can also be the unintended result of a number of variables, such as government demands, the type of task to be carried out, the vagueness or contradictory

nature of policy objectives, and the capabilities of the policy-implementing organization. Moreover, one is often not aware of the existence of a performance paradox until it is too late because as long as everything goes well—or appears to go well—there is no need to intervene (cf. Leeuw, 1995). A tragic example is found in the explosion of a firework factory, killing more than 20 people in the city of Enschede, the Netherlands, in May 2000. The investigation into the causes of this disaster revealed a range of "small" problems that by themselves were not considered to be catastrophic. For example, there was a clear lack of monitoring by the local and central governments, inspectorates, and the fire department. The absence of proper supervision prevented the discovery of illegal activities taking place. Hence, the license to operate in a residential area was—unjustly—renewed. When a fire occurred on the grounds of the factory, the neighborhood was destroyed and lives were lost. The accumulation of small problems had big consequences, but apparently there is no mechanism or system to detect and avoid accumulation of such small errors. Of course, local politicians were held accountable, but only after the fact. This brings us to the question of how one can detect—and prevent—the occurrence of a performance paradox.[10]

Several strategies are available to try to trace a performance paradox. Ideally, a comparison of reported and actual performance is the best way. However, such a comparison is generally very difficult to make because of the lack of comparative information. Alternative methods are (a) to use external sources to obtain information such as the national ombudsman, grassroots organizations, and client panels; (b) to develop new performance indicators from the existing indicators, for example, a percentage score instead of total expenditures or total output; or (c) to analyze the performance assessment system.

An analysis of the performance assessment system would have to focus on a number of characteristics. First, the number of indicators is important, as well as whether indicators have been developed for all tasks that have to be carried out. Few indicators for a limited part of total performance facilitate the occurrence of a performance paradox. This effect is reinforced when indicators do not change over time. Next, attention should be paid to the question of who develops the indicators. Organizations that develop their own indicators have more opportunities to manipulate information to their benefit and thus evoke a performance paradox (Van Thiel, 2001). However, when the principal is the only one who develops indicators, it runs another risk, namely, that it will only get the information that it requires. Executive agents will generally not volunteer to produce more information than requested, particularly if that is not in their own interest. Third, it needs to be established whether all accountability requirements are met and, if not, on which aspects of performance the agent does not report and why. Gaps in the performance reports could point to overrepresentation or underrepresentation of aspects of performance and thus to a performance paradox. Finally, the administrative and organizational underpinning of the performance assessment system should be investigated. When an organization is strongly guided by user manuals, lists of frequently asked questions, and procedures on how to handle a request from an auditor, the occurrence of a performance paradox increases.[11]

An analysis of the assessment system could be undertaken by comparing reports on performance over a number of years, preferably between organizations, by holding interviews with informants from both the executive agent and the government and by

asking for expert evaluations of the performance assessment system and its administrative underpinning (e.g., from consultants or accountants).

The analysis of the performance assessment system can reveal the presence of some of the aforementioned conditions that elicit a performance paradox. To prevent a performance paradox, Meyer and Gupta (1994) recommend that organizations adopt a so-called paradoxical model of performance assessment with multiple, uncorrelated, and varying but comparable performance indicators. They also recommend the use of targets and comparisons over time, between organizations, and/or between different units within the same organization.

Such a paradoxical model has to do a number of things. First, it has to find a balance between an expansion of the number of performance measures on one hand and a reduction of the measure pressure on the other hand. Both excessive and minimalist emphasis on performance indicators can result in a performance paradox. Second, as Bouckaert and Balk (1991) maintained, we should search for optimal measures that minimize the dysfunctional effects and maximize functional effects. A system of performance assessment in the public sector has to be able to cope with paradoxes and ambiguity. And third, it has to leave room for multiple interpretations of policy goals. Funders, purchasers, providers, and consumers have different interests in policy implementation, leading to different emphases in performance assessment.

Performance assessment in the public sector has to take the nature of public services into account (McGuire, 2001). "The way professional services are produced and consumed (delivered) and the way public services are valued by the community have implications for performance monitoring" (McGuire, 2001, p. 8). In the public sector, consumers participate in the service delivery process, affecting output and outcome (cf. Fountain, 2001, p. 58). Moreover, most products are intangible. Performance indicators should therefore reflect quality and reliability rather than "hard" product attributes. Public services are not only about efficiency and effectiveness but also about justice, fairness, equity, and accountability. Fountain (2001) warns that the application of private sector techniques, such as performance indicators, as propagated by NPM, cannot replace, indeed may obscure, such political or democratic outcomes of public service provision.

McGuire (2001) discusses an example of a performance-monitoring framework that seems to take into account some of the lessons discussed before. This framework was developed for the Council of Australian Governments (COAG) by the Productivity Commission to benchmark the performance in the education, health, housing, and community services (available on the Internet at www.pc.gov.au/gsp). The framework is developed in cooperation with all governments in Australia. It discusses the limitations of performance indicators at length, as well as the complexity of measuring human service provision. It includes both program and operational indicators, measuring both efficiency (output) and effectiveness (outcomes) along a number of different dimensions. Quantitative measures are combined with contextual analyses of the service systems.

The COAG performance information is used by government agencies to assess performance and determine needs and resources. The framework improves the transparency of performance and accountability. However, this transparency also "increases rather than resolves political conflict over the distributional consequences of provid-

ing public services" (McGuire, 2001, p. 17). Because political conflict increases the opportunities for accountability, it should not necessarily be considered to be a negative consequence of performance assessment.

Conclusion

The increase in performance assessment in the public sector following the administrative reforms of the 1980s and 1990s has had several unintended consequences, threatening insight into performance and performance itself. To counteract these consequences, performance assessment systems should take the special characteristics of the public sector into account. The contested nature of performance indicators requires the use of multiple indicators, referring to different aspects of policy implementation (tangible and nontangible) and reflecting the interests of all stakeholders (politicians, managers, funders, providers, purchasers, and consumers). Moreover, a balance has to be found between too much and not enough measure pressure.

The rise of a number of new monitoring mechanisms could prove helpful in the fight against unintended consequences like the performance paradox (cf. also Power, 1997). For instance, the Internet makes information on public sector performance accessible for everybody, increasing the risk for cheating executive agents to get caught. And secondly, Citizen Charters, the open government code of practice and new complaint procedures, increase the opportunities for unsatisfied customers to address ill-performing organizations. These new, more horizontal forms of performance evaluation will supplement performance assessment systems in the public sector.

Finally, academics should start to formulate and test theories that can explain the occurrence of a performance paradox and other perverse effects (see, e.g., Scott, 2001). More knowledge on organizational behavior and the influence of institutions and public sector characteristics on the use of performance indicators can help to truly achieve the projected advantages of performance indicators in the public domain.

Notes

1. In the case of the Netherlands, the number of review offices, supervisory organizations, inspectorates, and evaluation units has been enlarged in the past 10 years with a dozen or so new organizations, mainly regulators for privatized and autonomized services such as the police force and the telecommunications sector. Also, the budget for the Dutch National Audit Office has increased strongly since the early 1980s. Similar developments can be found in the United Kingdom (Power, 1997).

2. Dr. C. R. Bayer works at the Institut für Wirtschaftstheorie und OR, at the Universität Karlsruhe (TH), Germany.

3. Quasi-autonomous nongovernmental organizations (quangos) are organizations charged with policy implementation as their main task, paid by the government to do so but operating at arm's length without an immediate hierarchical relationship to that government (cf. Van Thiel, 2001). The British refer to nonexecutive public bodies or extragovernmental organizations (Hall & Weir, 1996). Well-known examples of quangos are semi-autonomous police authorities, universities, chambers of commerce, and social benefits agencies. Quangos can be found at national and local government levels.

4. Data were used from a contest on the best annual report for a public sector organization, organized by Arthur Andersen in the Netherlands between 1994 and 1999. Annual reports were evaluated on the presence or absence of certain performance indicators.

5. The number of dropouts could increase in two ways. First, the school might exert pressure to persuade ill-performing students to be absent when tests are administered. Second, students with learning problems or disabilities may not be able to cope with the test pressure and leave school entirely.

6. The principal-agent literature describes a comparable phenomenon: adverse selection, also known as the hidden information problem (Douma & Schreuder, 1998; Pratt & Zeckhauser, 1991). When drawing up a contract, how can the principal be sure that the agent can (and will) deliver what he or she promises? Monitoring is often mentioned as a solution to this problem. However, we argue that monitoring itself may have unintended consequences.

7. This problem resembles the problem of moral hazard in principal-agent theory (see also Note 6). After a contract has been signed, an agent may try to shirk and then use its information advantage to obscure performance measurement.

8. The competitive pressure for public sector organizations in so-called internal markets (i.e., competition for budgets between hospitals, schools, or bureaus of legal aid) can be counteracted by cooperation among the organizations, for example, through division of regional markets or cartel-like alliances (cf. Van Thiel, 2001).

9. In the principal-agent literature, this phenomenon is known as the reversal of control between agent and principal (White, 1991). It is no longer the principal who determines the contractual specifications but the agent due to his or her expert knowledge (information asymmetry) and the sunk costs of previous investments.

10. The Dutch Ministry of Finance is aware of the performance paradox and has cautioned other departments that—when reporting on their performance—they should know what constitutes "good performance" themselves rather than depend on policy-implementing organizations for this information (Van der Knaap, 2000). To avoid opportunism by executive agents, ministries were also advised to invest in a trusting relationship with executive agents by, for example, sanctioning a lack of performance information rather than ill performance. Also, agents should be given the opportunity to explain causes of failure, particularly when performance cannot be expressed in quantitative indicators.

11. John Meyer gave an example of this "orchestration" of audit during a speech at the European Union convention on the meaning of quality in education in 2001. He told about booklets and manuals that are available to head teachers in California that help them to fill in "all the requested formats on the performance and quality of schools in this state." In the words of Meyer, this went so far that "even the most unintelligent head teacher is able to produce minimal, but acceptable data."

References

Algemene Rekenkamer. (2000). *Organisatie van Beleidsevaluatie [Organization of policy evaluation] (Tweede Kamer, vergaderjaar 1999-2000, 27 065, No. 2). The Netherlands: Second Chamber of Parliament.*

Bouckaert, G., & Balk, W. (1991). Public productivity measurement: Diseases and cures. *Public Productivity & Management Review*, *15*(2), 229-235.

Bouckaert, G., & Ulens, W. (1998). *Prestatiemeting in de overheid: internationale lessen voor de Belgische overheid* [Performance measurement in the government: International lessons for the Belgian government]. Leuven, Belgium: Instituut voor de overheid, Catholic University Leuven.

Commissie Cohen. (1997). *Eindrapport Werkgroep Markt en Overheid* [Final report of the working committee market and government]. The Hague, the Netherlands: Ministerie van Economische Zaken, Commissie Marktwerking Deregulering en Wetgevingskwaliteit.

Douma, S., & Schreuder, H. (1998). *Economic approaches to organizations* (2nd ed.). New York: Prentice Hall.

Fitz-Gibbon, J. (1997). *The value added national project: Report to the secretary of state*. Durham, NC: University of Durham, School Curriculum and Assessment Authority.

Fountain, J. E. (2001). Paradoxes of public sector customer service. *Governance*, *14*, 55-73.

Ghoshal, S., & Moran, P. (1996). Bad for practices: A critique of the transaction cost theory. *Academy of Management Review*, *21*, 13-47.

Goodlad, S. (2000). Benchmarks and templates: Some notes and queries from a sceptic. In H. Smith et al. (Eds.), *Benchmarking and threshold standards in higher education*. London: Kogan Page.

Hall, W., & Weir, S. (1996). *The untouchables: Power and accountability in the quango state* [Democratic Audit of the United Kingdom]. London: Charter 88 Trust.

Heckman, J., Heinrich, C., & Smith, J. (1997). Assessing the performance of performance standards in public bureaucracies. *American Economic Review*, *87*(2), 389-395.

Heinrich, C. (1999). Do government bureaucrats make effective use of performance management information? *Journal of Public Administration Research and Theory*, *9*(3), 363-393.

Higher Education Funding Council for England. (2000). *Better accountability for higher education*. London: PA Consulting Group. Available from www.hefce.ac.uk.

Hood, C. (1994). A public management for all seasons. *Public Administration*, *69*, 3-19.

Jenkins, W. I., Leeuw, F. L., & Van Thiel, S. (in press). Quangos, evaluation and accountability in the collaborative state. In A. Gray (Ed.), *Evaluation and New Public Management*. London: Transaction.

Leeuw, F. L. (1995). Onbedoelde gevolgen van bestuurlijke intenties [Unintended consequences of policy intentions]. In H. van Gunsteren & E. van Ruyven (Eds.), *Bestuur in de ongekende samenleving* (pp. 55-72). The Hague, the Netherlands: Sdu.

Leeuw, F. L. (2000). Unintended side effects of auditing: The relationship between performance auditing and performance improvement and the role of trust. In W. Raub & J. Weesie (Eds.), *The management of durable relations*. Amsterdam: Thelathesis.

Leeuw, F. L., Toulemonde, J., & Brouwers, A. (1999). Evaluation activities in Europe: A quick scan of the market in 1998. *Evaluation*, *5*, 487-496.

Leeuw, F. L., & Van Gils, G. (1999). *Outputsturing in de publieke sector: een analyse van bestaand onderzoek* [Performance management in the public sector: An analysis of existing research]. The Hague, the Netherlands: Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

LeGrand, J. (1991). The theory of government failure. *British Journal of Political Science*, *21*, 423-442.

LeGrand, J., & Bartlett, W. (Eds.). (1993). *Quasi-markets and social policy.* London: Macmillan.

Light, P. C. (1993). *Monitoring government; inspectors general and the search for accountability.* Washington, DC: Brookings Institute.

McGuire, L. (2001, April). *Counting performance or performance that counts? Benchmarking government services in Australia*. Paper presented to the panel session on agencies and autonomous organizations at the Fifth International Research Symposium on Public Management, University of Barcelona, Spain.

Meyer, M. W., & Gupta, V. (1994). The performance paradox. *Research in Organizational Behavior*, *16*, 309-369.

Meyer, M. W., & O'Shaughnessy, K. (1993). Organizational design and the performance paradox. In R. Swedberg (Ed.), *Explorations in economic sociology* (pp. 249-278). New York: Russell Sage Foundation.

Mol, N. (1996). Performance indicators in the Dutch Department of Defence. *Financial Accountability & Management*, *12*(1), 71-81.

Organization for Economic Cooperation and Development. (1996). *Performance auditing and the modernisation of government*. Paris: Organization for Economic Cooperation and Development Press.

Osborne, P., & Gaebler, T. (1992). *Reinventing government: How the entrepreneurial spirit is transforming the public sector.* Reading, MA: Addison-Wesley.

Pollitt, C., & Bouckaert, G. (2000). *Public management reform: A comparative analysis*. Oxford, UK: Oxford University Press.

Power, M. (1994). *The audit explosion*. London: Demos.

Power, M. (1997). *The audit society.* Oxford, UK: Oxford University Press.

Power, M. (2000). The audit society—Second thoughts. *International Journal of Auditing*, *4*, 111-119.

Pratt, J. W., & Zeckhauser, R. J. (1991). *Principals and agents*. Boston: Harvard Business School Press.

Schmidtlein, F. A. (1999). Assumptions underlying performance budgeting. *Tertiary Education & Management*, *4*, 1-16.

Schwarz, R. (Ed.). (in press). *Assessing evaluative information: Prospects and perversities—Part II: Performance reports*. New Brunswick, NJ: Transaction Publishing.

Scott, W. R. (2001). *Institutions and organizations*. Thousand Oaks, CA: Sage.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, *18*, 277-310.

Torenvlied, R. (2000). *Political decisions and agency performance*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

[AUTHOR AVAILABLE?]. ***USA Today***. (2000, July 13). *Test pressure blamed for cheating*, **[PAGE?]**.

Van der Knaap, P. (2000). Resultaatgerichte verantwoordelijkheid: naar een beleidsgerichte begroting en verantwoording [Results-oriented responsibility: Towards a policy-based budget and accountability]. *Bestuurskunde*, *9*(5), 237-247.

Van Thiel, S. (2001). *Quangos: Trends, causes and consequences*. Aldershot, UK: Ashgate Publishing.

White, H. C. (1991). Agency as control. In J. W. Pratt & R. J. Zeckhauser (Eds.), *Principals and agents* (pp. 187-213). Boston: Harvard Business School Press.

Wiebrens, C., & Essers, S. (1999). Schaf het opheldering spercentage af [Abolish the percentage of solved crimes]. *Het Tijdschrift voor de Politie*, *61*, 27-34.

Wilson, J. Q. (1989). *Bureaucracy: What government agencies do and why they do it*. New York: Basic Books.

World Bank. (1998). *Assessing aid: What works, what doesn't and why*. Oxford, UK: Oxford University Press for the World Bank.

Dr. Sandra van Thiel is an assistant professor of public administration at Erasmus University Rotterdam, the Netherlands. Contact: P.O. Box 1738, 3000 DR Rotterdam, the Netherlands; fax: 31.10.4089099; e-mail: vanthiel@fsw.eur.nl

Dr. Frans L. Leeuw is director at the Dutch Educational Review Office and a part-time professor of sociology at Utrecht University, the Netherlands. Contact: fax: 31.30. 6666405; e-mail: flleeuw@cuci.nl