

PERFORMANCE MEASUREMENT: AN EFFECTIVE TOOL FOR GOVERNMENT ACCOUNTABILITY? THE DEBATE GOES ON

Patria de Lancer Julnes, Ph.D
Utah State University

Abstract

For much of the 20th century, accountability and performance measurement in the public sector centered on financial accounting, focusing on questions of how much money was spent and on what. Improved performance was mostly defined in terms of managerial efficiency. Recently, however, accountability has taken on a broader meaning to include the results of public actions. This emphasis on 'managing for results' has yielded the GPRA approach in the U.S. government. Efforts to promote accountability with this emphasis, however, have occasioned a backlash. In particular, some have criticized the information that results from performance measurement systems as inadequate for the task of guiding government resource allocation decisions. That task, say critics, is the domain of program evaluation. In reviewing the contributions of performance measurement and its limitations, I conclude that accountability needs are better addressed when program evaluators and performance measurement practitioners cooperate.

Running Head: de Lancer Julnes: Performance Measurement, Evaluation and Government Accountability.

Key Words: Accountability, performance measurement, program evaluation, implementation, adoption.

Patria de Lancer Julnes, Ph.D., is Associate Professor in the Political Science Department and Coordinator of the Masters of Social Science program in Public Administration. She is former co-Chair of the Center for Accountability and Performance of ASPA. Mail correspondence: 0725 Old Main Hill, Logan, Utah 84322; e-mail: patria.julnes@usu.edu

This paper has been accepted for publication in Evaluation by SAGE Publications Ltd. All rights reserved. ©Patria de Lancer Julnes

Accountability is a concept that has gained great notoriety, particularly during the past decade. Its meaning has been the subject of great deliberation in the United States, and translating the word into other languages has proven difficult. For example, at their annual congress two years ago, the Center for Latin American Development and Administration (CLAD), a Latin American association with its headquarters in Venezuela, could not agree on the appropriate Spanish word and decided to adopt the word “Accountability” in its English form. Of course, this was not sufficient to settle the debate as to the meaning of the concept. As a result, CLAD’s congress last year had two panels devoted to discussions on how to best characterize the meaning of the word in Spanish!

In the United States, evaluators have settled on oversight and compliance as the most accurate representation of the accountability concept (Mark, Henry and Julnes, 2000). Related to this operationalization, for the purpose of this paper we will understand the concept of accountability as holding someone responsible for something. Thus, being accountable implies responsibility for ones’ actions and their consequences (Roberts, 2002). This, of course, suggests a direct causal relationship between actions and results (Behn, 2003a), a point of contention for program evaluators.

The means for holding someone accountable is the domain of what Roberts (2002) calls performance-based accountability, which “requires the specification of outputs and outcomes in order to measure results and link them to goals that have been set, in accordance with the norms of management practice” (p. 659). Performance-based accountability requires performance measurement, the on-going production of information about an organization’s actual outputs and results as measured against its goals and objectives.

The current emphasis on results for both accountability and performance measurement has been associated with the increased skepticism and discontent of the American public with how their tax dollars are being spent. The apparent lack of government accomplishments spurred a movement demanding accountability in terms of results or outcomes. Performance measurement with an emphasis on outcomes has been heralded as one way to respond to these demands for results-oriented accountability. This has not gone unopposed. Among the many criticisms, some critics argue that performance measurement is an inadequate tool for government accountability for social spending. That, according to evaluators such as Greene (1999) and Ryan (2002), is the domain of program evaluation.

Given this background, in this paper I address three issues related to the use of performance measurement for the purpose of accountability:

- 1) Evolution of the meaning of accountability and performance measurement
- 2) Claimed contributions of performance measurement
- 3) Complementary forms of performance measurement and program evaluation.

I discuss the first issue from a historical perspective, starting with the work of Woodrow Wilson in the late 1800s. Then, I discuss current claims of purported benefits of performance measurement. In the last point I argue for a conciliatory stance between performance measurement and program evaluation.

Evolution of Accountability and Performance Measurement:

Toward A Broader Meaning of The Terms

The interest of the American public on accountability and performance measurement can be traced back to the writings of Woodrow Wilson in the latter part of the 19th century. Looking

at the pervasiveness of patronage and corruption, Wilson (1968) recognized the need to hold workers accountable for their actions. This, he believed, could only be accomplished through professionalism, which would, in turn, lead to honest and efficient government. The “progressive” movement that ensued viewed accountability as tantamount to government efficiency. Administration became viewed primarily as a technical question and the emphasis was on procedures and measurement techniques to identify and increase the productivity of workers (Bouckaert, 1992). Efficiency, narrowly defined as the ability to produce more output with less input, or the comparison of inputs to outputs (managerial efficiency), became the basis of scientific management studies.

The result of this emphasis on managerial efficiency was that accountability came to be understood primarily as financial accountability. And, performance measurement, so defined, came to be the tool for addressing accountability. Therefore, we see that early performance improvement efforts between the 1940s and 1970s, such as Planning Programming Budgeting System (PPBS), Management by Objective (MBO), and Zero-Based Budgeting (ZBB), had measurement at their core. Much has been written about the fate of these efforts, and thus I will not expound here. Suffice it to say, dissatisfaction with the performance of government programs, and particularly social programs, have continued to increase as well as citizens’ demand for evidence of program effectiveness (Wholey and Newcomer, 1997; Schoor, 1997). As a result, American public management has witnessed a move towards a broader understanding of efficiency, one that emphasizes effectiveness or the extent to which organizations are accomplishing their mission. For this, public agencies have started to pay attention to responding to clients’ needs (Miller, 1991).

The increasing lack of trust in government's abilities and the continued cynicism about the money spent by public organizations on social services also led to an emphasis on minimizing the role of government in delivering public services (Schein, 1996). As a result, in the 1980s privatization was promoted under the assumption that the private sector could do a better job than the public sector at delivering public services.

Along with privatization, and in response to the growing calls for accountability in the 1980s, state and local governments embraced the private sector's Total Quality Management (TQM) movement. Attractive features of this effort were its reliance on data collection and process analysis as well as a focus on responding to customers' needs (Hatry, 1997). Less attractive to some was the consequence of defining citizens as simply customers of public services.

In the 1990s the Governmental Accounting Standards Board (GASB),¹ began encouraging state and local governments to experiment with reporting service efforts and accomplishments (SEA), with *accomplishment* primarily concerned with results or outcomes of program activities. According to Hatry (1997), GASB can be credited with the increased efforts by state legislatures to require performance measurement by state agencies. These performance measurement efforts have focused on the identification of "outcome-oriented indicators for broad, statewide objectives, somewhat resembling the social indicators movement of the 1970s" (Hatry, 1997, p. 32). As a result, performance measurement systems are now expected to include the following types of indicators or measures:

- *Inputs* are resources used to produce outputs. These may include dollar costs, staff and staff time, materials, and other resources.
- *Outputs* are the final product or service delivered that ultimately, it is hoped, will lead to a desired outcome. Number of teachers attending a workshop on curriculum improvement is an example of an output.
- *Outcomes* are the consequences of outputs and are often more complex to measure.

- *Intermediate outcomes*: an outcome that is expected to lead to a desired end but is not an end in itself. An example of intermediate outcomes, following the example above, would be teachers actually improving the curriculum used in their classrooms after completing the curriculum improvement workshop.
- *End outcomes*: the end result that is sought-- such as the number of students mastering a subject.
- *Efficiency measures* indicate the ratio of output to input or outcome to input. This is also call unit-cost ratio.
- *Explanatory Information* provides the context for readers to interpret data. This is especially important when outcomes are poorer or better than expected. (Hatry 1999; de Lancer Julnes, 2003).

The evolution of accountability from accounting of inputs and outputs toward reporting of outcomes also happened at the federal level, culminating in 1993 with the passage of the Government Performance and Results Act, GPRA. This act, embraced by then Vice-President Gore's National Performance Review initiative, was designed to improve the effectiveness of federal programs and citizen satisfaction through the systematic measurement and reporting of performance measures that are aligned with agencies' goals. Under GPRA, federal agencies are required to develop quantitative performance measures and targets and to report these indicators to help monitor their performance (Berman, 1998; Kettl, 1994; de Lancer Julnes, 2003). As noted, this redirection, from activities and processes to a focus on outcomes or results, provides more emphasis on addressing citizens' needs (Heinrich, 2002).

This evolution has not been limited to the United States. For example, Hatry (1997) describes the practice in New Zealand and Australia of giving government agencies more flexibility and authority in exchange for accountability on what the agencies produce. In a recent study commissioned by the New Zealand Treasury, Petrie and Webber (2001, p. 5) concluded that the practice has led to "a more responsive, innovative public sector delivering better services." In addition, the authors report an increase in efficiency, improved financial accountability, and improved overall fiscal control.

Some caution is required when drawing lessons from other countries, and it is not clear that the experiences of Australia and New Zealand represent unqualified successes. For example, among the weaknesses of the New Zealand public sector management regime, Petrie and Webber report misalignment of outputs and outcomes. Furthermore, the authors cite Brian Easton's contention that the reforms have led to reductions in service delivery and a lack of consideration of equity and fairness in "setting public sector remuneration" (p. 9). However, the larger point of the internationalization of performance measurement, with various countries providing lessons for each other, remains noteworthy.

A final point regarding the evolution of performance measurement is the central role now accorded to citizens. Citizens are viewed as important players in shaping the quality and responsiveness of government programs in their community (Epstein, et al., 2000). When government seeks meaningful citizen participation, it increases perception of legitimacy and trust in government. Further, participation, concludes Petts (2001), supports institutional legitimacy and the bottom-up approach to decision making. This, in turn, increases the acceptance of the decisions and the chances of implementation while reducing the likelihood of vocal opposition (Callahan, 2001). In addition, some argue that engaging citizens in performance improvement efforts can lead to the long term sustainability of the performance measurement initiative (Weidner and Noss-Reavely, 1996). For performance measurement, citizens' participation not only implies that the appropriate values are represented, but also that more meaningful measures of performance, those that matter the most to citizens, the ones whose lives government is affecting, are more likely to be developed.

To summarize, the demand for accountability in the United States is understood in terms of the performance or results of actions, and, to some, good performance has come to mean

whether constituents are satisfied with the way tasks have been performed (Romzek and Dubnick, 1998). Because of this emphasis on results, current performance monitoring efforts are called Managing for Results. This latest management reform initiative attempts to link measures to program mission, to set performance targets and to report regularly on the achievement of target levels of performance (Newcomer, 1997). Governments now attempt to show their constituents what they are getting for their tax dollars, how efficiently and effectively their tax dollars are spent, and how expenditures benefit constituents' lives (Callahan and Holzer, 1999).

Assessing the Contributions of Performance Measurement

As developed in the discussion above, performance measurement has been at the core of management reforms to enhance accountability. Strictly defined, performance measurement is the regular and careful monitoring of program implementation and outcomes. This regularity is one of the characteristics of performance measurement that differentiates it from program evaluation. In what follows, I review claims of purported benefits of performance measurement and some of the concerns expressed by others. I then present findings of a study designed to help us understand how performance measurement information is actually used.

Claims and Concerns

Wholey and Newcomer (1997) have argued that performance measurement systems provide information that can be used to improve management and program effectiveness, improve policy decision making, and improve public confidence in government. This is accomplished by using performance measurement information for budget formulation and

resource allocation, employee motivation, setting goals and objectives, planning, evaluation, control, and for improving service quality and enhancing citizens' trust (Wang, 2002; Wholey and Newcomer, 1997; Behn, 2003b).

In discussing what performance measurement purports to contribute to the “contemporary clamor for results in our social programs,” Greene (1999, p. 161) expresses concern about four contributions often claimed by advocates of performance measurement: 1) agencies can easily make performance measurement information available to all audiences; 2) performance measurement systems are a “concrete instance of documented program outcomes” (p. 162); 3) because of their focus on outcomes and not compliance, performance measurement systems allow agencies to be innovative in their implementation of programs; 4) “a focus on outcomes provides communities (at all levels) with the opportunity for collective, shared deliberation about what constitutes valued outcomes from a given endeavor” (p. 162). Greene’s concerns about these claims may be a reflection of the fear expressed by Perrin (1998) that in some instances performance measurement information could be destructive because it may give the appearance of rational decision-making when in reality the information has been distorted to support political goals.

This fear is in stark contrast to the hopes of the earlier proponents of performance measurement and its predecessors (see Light, 1997, for a review of previous efforts) who envisioned a management information system in which decisions would be driven by the information collected. This optimistic view of rational decision-making and the pessimistic view of performance measurement as solely a political tool both presume instrumental use of performance measurement. However, as Weiss (1988) noted in her related work on the utilization of program evaluation, what we may really see is a less direct use of performance

measurement. In fact, performance measurement may have little direct impact on decision-making and still be of value in ‘enlightening’ various stakeholders. Current research reports an increasing number of public organizations saying that they have some type of performance measurement system (Berman and Wang, 2000; GASB, 1997). However, actual use of performance measurement information is not as common. On the contrary, research has shown that even when performance measures, and especially outcome measures, have been developed they often remain unused by public agencies (Behn, 2002; de Lancer Julnes and Holzer, 2001).

Empirical Elaboration

Evidence of underutilization of performance measurement information indicates a need to know more about the factors that affect utilization and the meaning of utilization. To that end, in 1997 we conducted a survey of a sample of 934 individuals (513 respondents, a 55 percent response rate) working for state, county, and municipal organizations in the United States. In brief, this survey, described in de Lancer Julnes and Holzer (2001), found that the utilization of performance measurement is composed of at least two stages, adoption and implementation. Adoption—understood in terms of a capacity to act-- refers to the development of measures of outputs, outcomes, and efficiency. Implementation—defined by the authors in terms of knowledge converted into action—refers to actual use of performance measures for strategic planning, resources allocation, program management, monitoring, evaluation, and reporting to internal management, elected officials, citizens or the media.

de Lancer Julnes and Holzer (2001) concluded that the two stages of utilization were differentially influenced by a number of factors. Adoption was more heavily influenced by rational/technocratic factors such as: the existence of an internal agency requirement to use

performance measures, availability of resources, a goal orientation in the organization, and available technical knowledge about performance measurement. Implementation, on the other hand, was more influenced by political/cultural factors such as external interest groups, the organization promoting risk-taking among employees and attitudes toward performance measurement.

This study helped to clarify the debate about the relative import of rational/technocratic and political factors (each is more important for one of the two stages). But it raised other questions, including just what was meant by implementation. That is, an adequate understanding of how performance measurement information is actually being used by public administrators was still missing. To shed some light on this issue, I conducted 18 follow-up telephone interviews in the fall of 2000 with individuals working in organizations that had participated in the 1997 mail survey. The informants held positions such as finance directors, city administrators, budget directors, and legislative auditors in state, county, and local governments. The interviews lasted anywhere from 20 to 40 minutes depending, in most cases, on whether the organization had developed systematic performance measurement systems. While the results of the follow-up study may be limited in generalizability, they provide a deeper understanding of what it means to implement or actually use performance measurement information for an important subset of government agencies.

To understand what implementation (actual use) of performance measures really meant, I asked an open-ended question about use and then went over a list of specific uses that were included in the 1997 survey. Responses to these specific uses are reported in Graph 1 below.

[Graph 1 about here]

As can be observed in the graph, resource allocation and monitoring of programs were the two categories of used most often cited by respondents (60% or 11 informants). Comparing the percentage of respondents stating that performance measures are used for resource allocation to those of a more recent study of cities in the United States conducted by Poister and Streib (2005), suggests that organizations that participated in my follow-up study were ahead of the utilization curve. Of those participating in Poister and Streib' study, only 48% of 225 said that performance data played an important role in resource allocation.

Using performance measurement for strategic planning and for reporting to elected officials was reported by 50% (9) of my informants. A little over 40% (8 informants) said that performance measures were used to report to internal management. Reporting to citizens, a purported priority of performance measurement, was the least mentioned use of performance measurement (33% or 6 informants). These patterns are consistent with Poister and Streib's (2005) findings. They report that only 56% of their 225 respondents stated that performance measures are used to track the implementation of projects or initiatives called for by the strategic plan; 48% report performance measures associated with the strategic plans to the city council and 35% report to the public on a regular basis.

When the discrete responses were matched to the open-ended question, the analysis indicated that when the informant reported implementing performance measures for resource allocation, this meant that the agency is using the information as part of the budgeting process. Sometimes the information is used by executives or legislators to ask questions regarding why a certain target was or was not achieved. One respondent stated that "they are used to analyze what's going on in the program and to bring it to the attention of the city council." Another said, "they are also used to explain, not necessarily for decision making." Interestingly, all

respondents reporting that performance measures were used for resource allocation indicated that they were rarely used to cut budgets in cases of poor performance.

In theory, strategic management is supposed to encourage clear accountability because by setting clear goals, objectives and measures, it allows managers, legislators, and other stakeholders to assess how the use of certain resources are contributing to intended results of a particular program activity or service (Koteen, 1997). Such use is not apparent from the answers to my interviews. According to informants, using performance measures for strategic planning really means that the measures are included in the organization's five-year plan as targets. Reporting to elected officials means that this is done through the budget document. Likewise, reporting to citizens and the media meant that the budget document was made available, often at the library and in some instances through the web.

To recap, the most notable common theme in the open-ended responses to my interviews was that performance measures do not drive decisions but are important and somehow influence action. To those who would judge the effectiveness of performance measures in terms of their direct use in decision-making, the conclusion would therefore be that performance measurement does not have an impact on program management. But this interpretation would miss the more subtle, but still valuable, impacts of performance measures. Specifically, respondents reported using performance measures (1) to justify budgets, (2) to fulfill mandates for transparency (accountability as oversight and compliance), and (3) to inform debate among administrators and elected officials.

Now, when it comes to the role of accountability, understood in terms of oversight, the results of the interview are paradoxical. Accountability emerged as a double-edged sword. On the one hand, the calls for accountability have encouraged some to use performance measures.

Eleven (61%) of the respondents mentioned accountability as the reason for using performance measurement information. On the other hand, accountability was also perceived as a deterrent to developing and implementing this type of information.

For example, while two respondents said that they would use performance measurement information for accountability purposes if pressured to do so, two others expressed concerns. In particular, one respondent said that the “management team is not receptive to performance measurement. There are concerns as to what the numbers say; how they present the issues...they are worried about accountability issues.” Likewise, another respondent said that “agencies are afraid of being held accountable for outcomes...Managers at all levels have fears that they set nice goals but can’t achieve them.” The fears about performance measurement and accountability, according to another respondent, are real. Performance measurement could open the door for “council to micro-manage and to set performance measurement standards arbitrarily.” For performance measurement efforts to be successful, suggested the respondent, there has to be agreement that performance measurement is a management tool “to help us change the way we do business to do our job better rather than a method meant to be an accountability hammer or to significantly change budget allocation.”

Furthermore, this last point was echoed by another respondent when commenting on the circumstances that complicate performance measurement in his organization. The respondent said that the perception of performance measurement in his organization is negative. The question in employees’ minds, said the respondent, is whether they will be punished if they measure and find that their program is not working. Thus, said the respondent, if you are going to promote performance measurement, you need to “emphasize that this is going to be used to help them improve.”

A related concern with performance measurement for accountability, one that should be of particular interest both to program evaluators and proponents of performance measurement, deals with how to develop a meaningful performance measurement system. Approximately 56% of respondents saw this as a challenge. They wondered what appropriate indicators of performance are, how to account for outcomes, what can be quantified, how to collect quality data, and how to report results. Some acknowledged that in their organization there was not a clear idea of what performance measurement is nor of how goals and objectives are developed and how to tie such things to budget decisions. Related to these concerns, another respondent said that even though his organization is required to “explain what we’ve done with the money,” when they started reporting performance data there was little guidance on how to do it. The respondent said that although it is getting easier, “there is not a lot out there that says how to measure.”

Taken together, responses to my telephone interviews suggest that performance measures have greater impacts on decision-making than some skeptics have concluded but also more subtle impacts than pessimists have feared and proponents have claimed. If we are able to use performance measurement to make more informed decisions based on a better understanding of a program, then we can say that performance measurement is useful. Likewise, even if the information is used in a political mode to justify our budget requests, there is still a contribution to transparency and accountability that too can be claimed as a success for performance measurement.

In addition, the responses provide some useful insights for those promoting accountability whether through program evaluation or performance measurement.

Accountability appears to cause fear, particularly when the perception is that information will be

used to punish. Thus, one of the challenges that proponents of performance measurement and program evaluation may face is convincing those who will be held accountable as well as those who are accountability holders that the information should first be used to improve performance.

Complementary Nature of Performance Measurement and Program Evaluation

Arguing that performance measurement can be used for program accountability and improvement, Wholey (1996) has stated that performance measurement can serve both formative and summative purposes just as well as one-shot program evaluation studies. He also states that for “many policymakers and managers, performance measurement *is* evaluation,” (p. 146, emphasis in original). These assertions worry program evaluators who perceive performance measurement as fraught with flaws. These worries arise from attempts to substitute performance measurement for program evaluation. As others have noted (Mark, Henry and Julnes, 2000) however, this is a false choice. We need to, and can, promote performance measurement in a way that complements program evaluation. I will focus here on three of the alleged limitations of performance measurement as evaluation (attribution, goal displacement, representation of quality) and then proceed to discuss how program evaluation can strengthen the performance measurement effort and vice versa.

Attribution

One of the limitations highlighted by critics is that performance measurement information by itself does little to establish causal linkages. That is, only rarely can causality, or as it has come to be known *attribution* –concluding that program outcomes are a direct result of activities--be established with any confidence by outcome data collected on a routine basis (Scheiver and Newcomer, 2001). The problem is that there are almost always competing

explanations that cannot be ruled out with only outcome indicators. This is the invidious problem associated with internal validity which, according to Shadish, Cook and Leviton (1990), is not readily understood by most managers who, as a result, end up assuming unwarranted causality. Also, these data are not likely to help us rule out spuriousness.

Moreover, Perrin (1998) contends that without program evaluation information about a presumed causality between program activities and results, performance measurement information is insufficient for making decisions such as budget allocations. Consider a school district where a significant investment has been made on teachers' training under the assumption that such training would improve students' performance as measured by district's average score on standardized tests. If after a period of time the school district's students still perform below expectation, the decision of administrators might be to redirect the budget allocations away from teachers' training efforts and even punish the teachers with salary cuts. This, following Perrin's logic, would not be the right decision because it is possible that the reason students are performing below expectations has nothing to do with teachers or teachers' training. The situation might be due to factors outside the control of program teachers or to faulty program logic. In that accountability presumes that actions are linked to results, performance measurement systems appear to be inadequate for addressing demands for accountability (Greene, 1999).

Goal Displacement

Another limitation of performance measurement is goal displacement. Goal displacement refers to individuals' tendency to purposely change their behavior in the areas that they are being measured so to improve the performance ratings (Mark, Henry, and Julnes, 2000; Bohte and Meier, 2000). Unfortunately and to the detriment of the program, focusing on

improving the wrong behavior can happen at the expense of the more desirable program outcomes. The result is that not only are more important activities neglected and the program's clients with the most need disenfranchised, but also the validity of the particular measure is distorted. Sometimes this displacement is the result of reasonable efforts to address stated priorities. Other times, however, individuals know that they are forsaking the organization's larger interest and engage in what some consider cheating (Bohte and Meier, 2000).

Mark, Henry, and Julnes (2000), illustrate goal displacement with an example about teachers who start "teaching to the test" when they are rewarded based on their student's performance on standardized tests" (p. 204). In this instance goal displacement occurs if the teacher disregards other aspects of the desired learning experience and focuses on teaching students only the topics covered in the test. Taken further, the "goal" of improving test scores can lead to helping students answer the questions during the test, clearly undesirable results.

In sum, with goal displacement, use of the standardized test as a measure of "teacher effectiveness" is even less valid because what is being measured is not the overall effectiveness of the teacher in the classroom environment but rather his/her ability to teach students how to take the test. Similarly, we are no longer gauging students' learning; what is being gauged is students' ability to answer this particular test.

Capturing and Representing Program Quality

The third limitation to be addressed here is Greene's (1999) assertion that performance measurement does not adequately capture and represent program quality. With regards to capturing quality, Greene (1999) argues that human experience is too complex to be meaningfully reduced to a simple measure of program quality. Likewise, because the meaning of program quality varies from stakeholder to stakeholder and standards for judgments of

program quality are not objective and fixed, performance measurement systems cannot capture the critical dimensions of program quality.

When it comes to representation, measuring is not just about gathering the information but also “‘writing it up’ ... re-presenting what has been learned” (Greene, 1999, p. 166). Thus, Greene concludes that given the dynamic, pluralistic and contextual aspect of program quality, performance measurement systems can only represent a very small portion of what is important about a human experience. Problems with representation are important as value judgments in the real world almost always involve a balance among multiple, deeply-held values (Julnes and Foster, 2001).

These limitations have prompted program evaluators to go as far as calling for an abandonment of the practice of performance measurement and for evaluators to become more involved in the development of accountability systems. In the area of education accountability, for example, Ryan (2002) has suggested that because of the “collective wisdom residing within the evaluation community and its practice” program “evaluators are uniquely situated to make a significant contribution in the dialogue about the merits and shortcomings of educational accountability” (p. 454). Others, more supportive of performance measurement efforts (Wholey, 1996; Scheiver and Newcomer, 2001), have called for cooperative efforts between program evaluators and program managers, policy analysts and others to help at the various stages of developing and implementing a performance measurement system. Evaluators, states Wholey (1996), “should help us learn how to better use performance measurement data to improve government performance and credibility” (p. 148).

Overcoming limitations

Given the limitations discussed above, the question then is whether performance measurement is worth evaluators' and decision-makers' trouble. Available evidence suggests that answer is a clear 'yes.' Program evaluators who have tested the utility of performance measurement systems not only have concluded that they are useful but also that they can provide the backbone to more involved program evaluation efforts. For example, Harkreader and Henry (2000) stated that while performance measurement systems should be used cautiously to ascertain the worth of programs (value toward the larger social good), they are nonetheless effective tools for assessing merits (quality of program or policy in terms of performance) of programs. Furthermore, continued these authors, "we believe that evaluators can use performance measures as a low-cost way of providing indications of merits of a program" (p.167).

Another point to consider is the relative costs of performance measurement and program evaluation methodologies. On the one hand, research suggests that the cost and time that is required to fully develop and implement a performance measurement system are often cited as major barriers to performance measurement (Ho, 2003). On the other hand, properly conducted program evaluations are generally more expensive and more demanding of other scarce resources. As a result, program evaluation studies tend to be one-shot efforts. Furthermore, because program evaluations are episodic, the information that results is often unavailable at critical decision points and sometimes even out of date by the time that it is available. As such, the findings of program evaluations often go unused because decision makers find them untimely and irrelevant (Wholey, 1996). These points argue for the value of widespread use of performance measurement and more selective use of program evaluation.

Also, while not trying to minimize the valuable contributions that program evaluation can make, one can argue that there are instances when performance measurement information is sufficient to accept apparent causal relationships (i.e., the program is leading to good outcomes). According to Perrin (1998), this is particularly the case for discrete and relatively easy to deliver public services such as street cleaning and road maintenance. Threats to internal validity are not as critical for two reasons: 1) few other explanations exist (what else besides public services would have led to cleaner or repaired streets?); and 2) faulty conclusions can be easily detected and remedied. In such circumstances there would be no need to engage in a program evaluation study.

However, there are other instances where a more in-depth analysis is necessary, particularly when a program's outputs and outcomes are not up to expectations. The reason for this is that performance measurements do not tell us the determinants of the outcome, and so do not address questions of how and why. Yet, even in those instances, depending on what type of services we are talking about, simple techniques such as talking to citizens could help determine the reasons and not necessitate a full blown evaluation study. Furthermore, the concurrent monitoring of program implementation could help illuminate the reasons a program is not achieving desired results.

In addition, in that program evaluators are trained in a variety of methods and approaches, including logic models and different ways to collect, maintain and analyze data, their knowledge will prove useful for selecting and collecting multiple measures of a particular indicator and conducting more thorough causal analysis. Also, the traditions of program evaluation with their emphasis on construct validity can also help correct for goal displacement

because they call for a “broad array of performance indicators so that people would be unlikely to tilt their behavior toward a single indicator” (Mark, Henry & Julnes, 2000, p. 204).

At the same time, it is important to recognize that “every evaluation method, including PM [performance measurement], has its limitations which can only be overcome through use of a combination of methods” (Perrin, 1998, p.374). Thus, much of the criticisms presented here about performance measurement can also apply to program evaluation studies. Program evaluations need to rely on multiple sources of information; performance information, as suggested by Perrin (1998) and others (e.g., Harkreader and Henry, 2000; Scheiver and Newcomer, 2001), can be a key component of this needed multiplicity.

On the other hand, program evaluation can improve reliability for the performance measurement system. As demonstrated in the experience of the Australian public sector, while the more regular measurement of program performance can be an effective management and accountability tool, external program evaluation can provide verification of the performance measurement system and improve its credibility (Guthrie and English, 1997). Finally, indicators of performance that have been fashioned for a program evaluation would be more useful than untested measures of performance.

CONCLUSION

Given the long history of calls for accountability and performance measurement as a tool for answering those calls, it is reasonable to argue that performance measurement will continue to play a central role in addressing such calls. But the assumption of accountability, namely that an activity can be linked to a specific result, poses some challenges for performance measurement. Three of these challenges were discussed, attribution of results, rewarding

behavior, and measurement of multi-dimensional phenomena like program quality. As we reflect of these challenges, three concluding points emerge.

First, accountability, whether through program evaluation or performance measurement, will encounter the same challenge--fear, justified or not, on the part of those being held accountable. The main goal of accountability should be to improve service delivery, and not to punish in instances when services and programs are not performing up to expectations. Unfortunately, a related challenge will be convincing accountability holders of this.

Second, performance measurement can best be used for accountability purposes, defined in terms of oversight and compliance, under the following conditions: 1) the program logic for the service in question is not faulty. The lack of credible threats to internal validity warrants the assumption of attribution; 2) appropriate goals and objectives have been developed, thus the information obtained can be used to assess whether or not the program is moving in the desired direction; and 3) the service is not complex. In this last condition, the service delivered is simple enough that routine monitoring of activities or simple evaluation techniques can be used to assess outcomes. When the service is more complex, the guidance provided by performance measurement can serve not only to point out where more in-depth evaluation is needed, but, as argued by Harkreader and Henry (2000), also can serve as the backbone of evaluation studies; 4) the measures have been validated by a program evaluation.

Thirdly, as suggested above, the knowledge and skills of program evaluators can help to improve the quality of performance measurement systems. This need for knowledge and skills was further illustrated by some of the concerns expressed by respondents to the interviews reported here. Therefore, in order to support accountability and advance practice and theory,

what is needed is cooperation and not competition between those in the performance measurement and the program evaluation camps.

Notes

¹ GASB is a nonprofit organization that develops accounting standards for state and local governments for the purposes of external reporting.

REFERENCES

- Behn, R. (2003a). 'Rethinking Accountability in Education; How Should Who Hold Whom Accountable For What?', *International Public Management Journal* 6 (1): 43-73.
- _____. (2003b). 'Why Measure Performance? Different Purposes Require Different Measures', *Public Administration Review* 63 (5): 586-606.
- _____. (2002). 'The Psychological Barriers To Performance Management: Or Why Isn't Everyone Jumping On the Performance-Management Bandwagon?', *Public Performance and Management Review* 26 (1): 5-25.
- Berman, E. M. (1998). '*Productivity In Public And Nonprofit Organizations. Strategies And Techniques*', Thousand Oaks: Sage.
- Berman, E. M. and Wang, X. (2000). 'Performance Measurement in U.S. Counties: Capacity for Reform', *Public Administration Review* 60 (5): 409-420
- Bouckaert, G. (1992). 'Public Productivity in Retrospective', in M. Holzer (ed.) *Public Productivity Handbook*, pp: 5- 46. Marcel Dekker: New York.
- Bohte, J. and Meier, K (2000). 'Goal Displacement: Assessing The Motivation For

- Organizational Cheating', *Public Administration Review* 60 (2):173-182.
- Callahan, K. and Holzer, M. (1999). 'Results-Oriented Government: Citizen Involvement in Performance Measurement', in A. Halachmi (ed.) *Performance & Quality Measurement in Governmen. Issues and Experiences*, pp: 51-64. Burke, VA: Chatelaine Press.
- Callahan, K. (2001). 'Results-oriented government: Citizen Involvement in Performance Measurement', Paper presented at the Winelands Conference, Stellenbosch, South Africa: 12 September.
- de Lancer Julnes, P. (2003). 'Performance Measurement', in Rabin, J. (ed.) *Encyclopedia of Public Administration and Public Policy*, pp: 901-905. New York: Marcel Dekker.
- de Lancer Julnes, P. and Holzer, M. (2001). 'Promoting the Utilization of Performance Measures in Public Organizations: An Empirical study of Factors Affecting Adoption and Implementation', *Public Administration Review* 61(6): 693-708.
- Epstein, P., Wray, L., Marshall, M. and Grifel, S. (2000). 'Engaging Citizens in Achieving Results that matter: A Model for Effective 21rst Century Governance', Paper presented at the Center for Accountability and Performance's Symposium on Leadership of Results-Oriented Management in Government, Washington, DC: 11 February.
- Governmental Accounting Standards Board, GASB (1997). 'Report On Survey of State and Local Government Use and Reporting of Performance Measures: First Questionnaire Results'. Available at: http://www.seagov.org/sea_gasb_project/1997_Survey.pdf (site visited: 01 May 2005).

- Greene, J (1999). 'The Inequality of Performance Measurements', *Evaluation* 5 (2): 160-172.
- Guthrie, J. and English, L. (1997). 'Performance Information and Programme Evaluation in the Australian Public Sector', *The International Journal Of Public Sector Management* 10 (3): 154-164.
- Harkreader, S. A. and Henry, G. T. (2000). 'Using Performance Measurement Systems for Assessing the Merit and Worth of Reforms', *The American Journal of Evaluation* 21 (2):151-170.
- Hatry, H. (1999). *Performance Measurement. Getting Result*. Washington, DC: The Urban Institute.
- ___ . (1997). Where The Rubber Meets The Road: Performance Measurement For State And Local Public Agencies. *New Directions for Evaluation* 75:31-45.
- Heinrich, C. J. (2002). 'Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness', *Public Administration Review* 62 (6):712-725.
- Ho, A. (2003). 'Perceptions of Performance Measurement and the Practice of Performance Reporting by Small Cities', *State and Local Government Review* 35 (2): 161-73.
- Julnes, G. and Foster, E. M. (2001). 'Crafting evaluation in support of welfare reform', in G. Julnes and E. M. Foster (eds.) *Outcomes of Welfare Reform for Families Who Leave Welfare*, pp: 2-8, *New Directions for Evaluation*, 91.
- Kettl, D. (1994). *Reinventing Government? Appraising the National Performance Review*. Washington, DC: Brookings Institution.

- Koteen, J. (1997). *Strategic management in public and nonprofit organizations* (2nd edition). Westport, Connecticut: Praeger.
- Light, P. (1997). *The Tides of Reform: Making Government Work 1945-1995*. New London, CT: Yale University Press.
- Mark, M., Henry, G., and Julnes, G. (2000). *Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs*. San Francisco: Jossey-Bass.
- Miller, G. (1991). *Government Financial Management Theory*. New York: Marcel Dekker.
- Meld, M. (1974). 'The Politics of Evaluation of Social Programs', *Social Work* 19: 448-445.
- Newcomer, K. (1997). 'Using Performance Measurement to Improve Programs', *New Directions for Evaluation* 75:5-14.
- Perrin, B (1998). "Effective Use and Misuse of Performance measurement". *American Journal of Evaluation* 19 (3): 367-380.
- Petrie, M. and Webber, D. (2001). 'Review of Evidence on Broad Outcome of Public Sector Management Regime', Treasury Working Paper 01/06. New Zealand Treasury.
Available at: <http://www.treasury.govt.nz/workingpapers/2001/twp01-6.pdf> (site visited: 28 April, 2005).
- Petts, J. (2001). 'Evaluating the Effectiveness of Deliberative Process: Waste Management Case-Studies', *Journal of Environmental Planning and Management* 44 (2): 207-226.
- Poister, T. and Streib, G. (2005). 'Elements of Strategic Planning and Management in Municipal Government: Status after Two Decades', *Public Administration Review* 65 (1): 45-56.

- Roberts, N. C. (2002). 'Keeping Public Officials Accountable Thought Dialogue: Resolving the Accountability Paradox', *Public Administration Review* 62 (6):658-669.
- Romzek, B. and Dubnick, M. (1998). 'Accountability', in J. M. Shafritz (ed) *International Encyclopedia of Public Policy and Administration*, pp 6-11. New York: Westview Press.
- Ryan, K. (2002). 'Shaping Educational Accountability Systems', *American Journal of Evaluation* 23 (4): 453-468.
- Shadish, W., Cook, T, and Leviton, L (1990). *Foundations of Program Evaluation*. Thousands Oaks, CA: Sage.
- Schein, E. H (1996). 'Culture: The Missing Concept in Organization Studies', *Administrative Science Quarterly* 41: 229-240.
- Scheiver, M. and Newcomer, K. (2001). 'Opportunities for Program Evaluators to Facilitate Performance-Based Management', *Evaluation and Program Planning* 24: 63-71.
- Schoor, L (1997). *Common Purpose: Strengthening Families and Neighborhoods To Rebuild America*. New York: Anchor Books.
- Weidner, M. and Noss-Reavely, M. (1996). 'The Council on Human Investment: Performance Governance in Iowa', Washington, DC: ASPA's Center for Accountability and Performance (CAP).
- Weiss, C. H. (1988). 'Evaluation for decision: Is anybody there? Does anybody care?', *Evaluation Practice*, 9(1), 5-20.
- Wholey, J. (1996). "Formative and Summative Evaluation: Related Issues in

Performance Measurement”, *Evaluation Practice* 17 (2): 145-149.

Wholey, J. and Newcomer, K. (1997). ‘Clarifying Goals, Reporting Results’, *New Directions for Evaluation* 75: 91-98.

Wilson, W (1968). ‘Study of Administration’. *The Papers of Woodrow Wilson 1885-1888*. Princeton: Princeton University Press.

Graph 1. Distribution of Responses

