

**Contextual Approaches to Using
Measurement to Improve Performance**

George Julnes

**A Performing Public Sector:
The Second Transatlantic Dialogue**

June 1-3 2006

Leuven, Belgium

George Julnes

Department of Psychology

Utah State University

gjulnes@cc.usu.edu

435-797-1633

Contextual Approaches to Using Measurement to Improve Performance

The past half-century has witnessed increasingly sophisticated efforts in the enterprise to gather, interpret, and use information to improve the operations of the public sector. Performance measurement, a subfield of the larger enterprise of evaluation, has been a central component of this increasing sophistication and warrants our efforts to make it even more effective as an approach to improving performance. This paper supports these efforts by connecting performance measurement to other developments within the field of evaluation, emphasizing the contextual approaches to evaluation and their implications for performance measurement.

A. Performance Measurement and the Enterprise of Evaluation

We can begin our efforts by considering the efforts to recognize contextual factors in performance measurement and evaluation. For example, there is general consensus that performance measurement is used in pursuit of multiple objectives. Whereas many have promoted performance measurement as a tool for allocating resources (Newcomer, 1997), some have emphasized the role of performance measurement in increasing transparency in government (de Lancer Julnes, in press), and others seek to strengthen the related objective of citizen involvement in governance (Kelly, 2002; Kelly and Swindell, 2002; Dusenbury, Liner, and Vinson, 2000). Common among these various uses is the belief that collecting key information and providing it to public policymakers, administrators, and other stakeholders can improve decision-making in the public arena. This is a belief that underlies all of the subfields of evaluation.

Yet, the performance measurement field has no shortage of literature detailing the low utilization of available information by decision-makers (Berman & Wang, 2000; de Lancer Julnes and Holzer, 2001, de Lancer Julnes and Mixcoatl, in press). Of the wide variety of explanations offered for low utilization (e.g., the reluctance of policymakers to subordinate their autonomy in political process to supposedly rational data), one of the most cited reasons in the evaluation literature for a similar record of under-utilization is the lack of fit between the information provided and the information needs of the primary decision-makers. Patton (1997) has documented this lack of fit and has developed an approach, “utilization focused” evaluation,

that seeks to improve fit by focusing on the specific information needs of identified decision makers. This focus is behind Patton's adage of directing evaluation to the "intended use by intended users" (1997, p. 20). Intended users varies by situation but generally refers to people such as administrators who arrange for performance measurement or some other evaluative task because they believe it will provide the information needed in their setting. Similarly, Wholey (1979, 1983) emphasizes the importance of providing needed information with his notion of the "sequential purchase" of information. In this approach, there is a hierarchy of information that decision makers may wish to "purchase" in the sense of commissioning progressively more intensive inquiry approaches as the situation requires, from rapid feedback evaluation, to performance monitoring, to intensive evaluation.

This concern with the information needs of intended users resonates with some current concerns in the performance measurement field. For example, Halachmi (2002) highlights the difference between using performance measurement for the purpose of accountability and using it for the purpose of improving program and agency performance. Further, he adds a dynamic of professional orientation, claiming that auditors tend to use performance measurement for oversight and not for the improvement goals of those supposedly using the performance information. This suggests a need to be clearer about the various "purposes" of performance measurement. However, the conceptual parsing of purposes is not enough, we also need clear guidance on how to direct our measurement efforts in support of different purposes. Specifically, efforts to match information needs to the information-gathering activities work best when there are practical frameworks that help both policymakers and analysts. Thus, this paper develops a practical framework that begins with an analysis of purposes in the field of evaluation. We can then consider the implications for the choices that follow for performance measurement.

Describing this framework in a short paper will require some discipline in being both systematic and selective. Figure 1 presents the systematic logic for the material presented in this paper. To provide a brief overview of what is to follow, this system posits that there are three instrumental purposes of evaluation, including performance measurement, that are differentially important in different contexts—program oversight, program improvement, and resource

allocation. Along with these three instrumental purposes, there is also a more abstract knowledge development purpose, sometimes referred to as enlightenment or sensemaking, that is always important in performance measurement but is less frequently the stated purpose. Once the key players are able to settle on the most important, or primary (and perhaps secondary), purpose for a performance measurement system, we can design that system to accomplish the traditional evaluation tasks (implementation evaluation, outcome monitoring, and impact assessment) that are best suited to achieving that purpose (as indicated by the arrows in Figure 1 from components to purposes). Finally, having selected the appropriate evaluation tasks for a given situation, we can then choose the specific inquiry methods with which to accomplish those tasks, though this last step is beyond what will be addressed in this paper.

[Figure 1 about here]

B. Purposes of Evaluation

Addressing the information needs of the intended users of evaluations, including performance measurement efforts, requires first addressing who should be the intended users of an evaluation. The most general sense of who might be interested in the results of evaluations is conveyed by the term “stakeholder.” Stakeholders are understood in evaluation as referring to all people who have a significant interest, or stake, in the operation or outcomes of a program. Not surprisingly, there is considerable controversy over how inclusive to be in delineating who has a “significant” stake in a program, but most would agree that policymakers, program personnel, program users, and program funders all could be key stakeholders. If we are to design and conduct evaluations that address the needs of stakeholders, we need to understand these needs. Part of this challenge is in being open to what stakeholders are saying, but another part is being able to take an often cacophonous set of stakeholder questions and frame them in a way that leads to a coherent and responsive evaluation. The first part of this section presents the range of questions of interest to stakeholder; we can then consider a framework that organizes the many questions into a smaller set of four general purposes of evaluation. This four-purpose framework is then developed to help us choose the most important evaluation purposes for given situations.

Range of Stakeholder Questions

To help us appreciate the range of expectations confronting evaluators, Weiss (1998, p. 6) lists some of the major questions that administrators and others need help in addressing:

Many people want (and need) to know: How is the program being conducted? What is it actually doing? How well is it following the guidelines that were originally set? What kinds of outcomes is it producing? How well is it meeting the purposes for which it was established? Is it worth the money it costs? Should it be continued, expanded, cut back, or abandoned? Should people flock to it, sign up selectively, or stay away in droves? Does it work for everybody or only some kinds of people? (p. 6).

Several points stand out in this quote from Weiss (1998). First, note the diversity of these questions. There is no reason to expect the same types of evaluation designs and methods to be appropriate for all, or even for many, of these questions. The same contextual perspective is necessary for performance measurement, as well. For example, the use of performance measurement for assessing the fidelity of program activities with the original guidelines is quite different than using performance measurement to support reallocating resources from one program to another.

The second point to emphasize is that Weiss is not restricting these questions to administrators; as noted above, many people are ‘stakeholders’ in the sense of having a stake in the running of public programs, and most have some need for information. Because different stakeholders often are most concerned with different questions, evaluators generally experience an inherent tension in trying to provide information relevant to the full range of questions that are relevant to at least some stakeholders in a given situation. The basis of this tension is that (1) it is generally impossible to address all relevant questions in a given situation and (2) giving too high of a priority to the questions of some stakeholders can reduce the value of the evaluation for other stakeholders. This is important for performance measurement in that, as it matures as an

approach, there is increasing recognition of roles of citizens and other non-managerial people as stakeholders.

Framework of Primary and Secondary Purposes

Given the range of questions that can be relevant to different stakeholders and at different times in project life cycle, evaluators need some way of making sense of and organizing the many practical questions in a manner that helps us focus our activities for best effect. Scriven (1967) provided an initial distinction of two types of evaluation that differed according to purpose, with formative evaluation concerned with the purpose of promoting program improvement and summative evaluation focused on informing resource allocation decisions in selecting between or among program alternatives (this, for Scriven, is accomplished when evaluators provide comparative overall judgments of merit and worth). The idea behind this distinction is that these two types of evaluation, formative and summative, are directed to different purposes, and so you can focus your efforts more effectively by being clear which purpose is called for and then conducting the corresponding type of evaluation.

These two distinct purposes of evaluation were later augmented by a third, conducting evaluation with the purpose of yielding knowledge in service of better understanding or “enlightenment” (Patton, 1997; Weiss, 1998), or what we will call knowledge development. This emphasis on knowledge development was a consequence of a growing recognition that evaluations often did not affect specific decisions but rather had a more diffuse effect in framing how people looked at an issue. Related, organizational theorists have noted how knowledge about potential solutions can be filed away for years before the right combination of factors leads it to be connected to a problem and acted upon (e.g., see Cohen, March & Olsen, 1972) for a description of their “garbage can” model of bounded rationality). This is of particular interest for performance measurement in that administrators often report using the data that result in just this way, not to make decisions but to provide a frame for them. Thus, the “under-utilization” of performance measures may be a function of the type of use that is expected.

The last purpose generally recognized in evaluation is a direct function of the recurring emphasis on accountability in government and recent work in performance measurement (see de

Lancer Julnes, in press). This purpose, including both process and outcome monitoring, is referred to as program oversight (Mark, Henry, & Julnes, 2000; Wholey, 1997). While it is the purpose most closely connected with the performance measurement movement, the claim made here is that performance measurement has become broader than that and, at times, is used to address the other three purposes as well.

Further elaboration will no doubt lead to other purposes of evaluation being distinguished, but a framework based on these four purposes has much to offer. For example, a four-purpose framework makes it easier to address a primary and a secondary purpose without attempting to implement an all-purpose evaluation. Sometimes this happens when the relevant stakeholders agree that, for example, program oversight and improvement are paramount. Other times different stakeholder groups have different questions to be addressed, and the evaluation needs to be responsive to these differing needs (e.g., federal agency providing funding is concerned first with program oversight while the state agency running the program is interested in program improvement). In either of these situations, it is usually best for the sake of coherence and focus to try to identify a primary purpose of the evaluation but also be open to the need for a secondary purpose to be addressed. The notion of forcing a choice is that one rarely has the resources or time for an evaluation to address all purposes adequately.

The distinction between primary and secondary purposes is useful also in helping us appreciate the role of knowledge development as a purpose of evaluation. Whereas knowledge development, or enlightenment, can turn out to be the most important outcome of an evaluation, it is rarely if ever the focus of stakeholder questions. Rather, while knowledge development is always a consequence of good evaluation, most evaluation-related knowledge is intended as instrumental in ultimately supporting one or two of the three primary purposes. Thus, as depicted in Figure 1, any of the four purposes can be a secondary purpose of an evaluation, but only the instrumental purposes of program oversight, program improvement, and resource allocation tend to be presented as the primary purpose when designing and implementing evaluations, including those focused on performance measures. Accordingly, while acknowledging the centrality of knowledge development in all evaluations, we will focus our attention on the three instrumental purposes.

While each of these three purposes can be relevant for any stakeholder in certain contexts, there are some areas of traditional concern to different groups (see Rossi et al. 2004, pp. 34-37; Weiss, 1998, pp. 25-28; and Mark et al. 2000, pp. 54-58). To highlight these traditional areas of interest, Table 1 lists the purposes that are often most central for three stakeholder groups, policymakers, program managers, and what we can group together for convenience as consumers/citizens.

Table 1: Framework for instrumental purposes and concerns addressed by evaluation

	Policy Makers	Managers & Staff	Consumers
<p>Concerns Related to Program Oversight</p> <p>(1) Judging the adequacy of program use of resources</p> <p>(2) Judging the adequacy of program-related outcomes</p>	Central Concern		
<p>Concerns about Program Improvement</p> <p>(1) Making midcourse corrections</p> <p>(2) Studying program mechanisms for improvements</p>		Central Concern	
<p>Concerns for Selection Among Alternatives</p> <p>(1) Making continuation, expansion, or institutionalization decisions about a single program</p> <p>(2) Choosing the best of several alternative programs</p>	Central concern		Central concern

Policymakers are distinguished from program managers by their distance from program operations, though the distinction is not absolute. In general, policymakers include senior

administrators and legislators who are removed from the details of program operations but must make funding decisions about the programs. Program managers are likely to have little say about the funding of their programs (and little interest in advocating cancellation) but often considerable latitude in making incremental changes. Consumers include program participants (e.g., patients of a hospital) and other people who make decisions about participation (such as parents of participants or employers who fund employee training programs); citizens form a special category of people who pay the taxes that fund programs but can also play a more active role in guiding governance than the consumer model might suggest (de Lancer Julnes, 2006).

Program Oversight. The first set of concerns in Table 1, under the purpose of oversight, addresses two issues, first, the degree to which the program is operating properly and, second, the degree to which stakeholder needs are being met. As developed below, these two issues correspond to the two evaluation types shown in Figure 1 as serving program oversight, implementation evaluation and outcome monitoring. The important point about oversight is that the descriptive methods used have the goal not of solving problems but of identifying them so that greater attention can be given to them by policymakers and administrators. As such, oversight requires descriptions of program operations and outcomes and also criteria to determine what constitutes a problem warranting additional attention.

With regard to proper operation, oversight authorities such as legislatures are concerned with whether the program is following the original guidelines and is meeting the purposes for which it was established. The descriptive element is having detailed information about program resources used and activities conducted; the judgment comes in comparing the descriptive account to some criteria. For implementation, the criteria are often the specifications in the original guidelines, but activities are also compared to least implicit notions of best practice. With regard to meeting stakeholder needs, the oversight function involves developing a descriptive account of the degree to which needs are being met and unmet and comparing this to some criteria, either the degree of unmet needs at some other time or place or a socially defined level considered acceptable.

Managers are concerned with program oversight when they use a performance measurement approach for identifying areas needing greater attention. For example, measuring

citizen satisfaction with various outcomes can alert managers to areas needing to be addressed, particularly if there are comparisons with satisfaction at other times or in other communities. Consumer involvement in oversight is more recent, but many projects have advisory boards with substantial consumer representation. In addition, the move to engage citizens in pursuit of this purpose is gaining strength with citizen-driven performance measurement systems (de Lancer Julnes, in press). In sum, common to all of these forms of program oversight is the use of evaluation information to highlight issues needing greater attention.

Program Improvement. The second set, concerned with the purpose of making changes that lead to program improvement, is the focus of what is generally called formative evaluation. The most common exemplars of program improvement involve program managers and staff trying to make changes, often incremental, that lead to better implementation or more effective program delivery. More is demanded of methods used for this purpose than for oversight in that the goal is not merely to direct attention but rather to guide change. Such guidance requires enough evidence to warrant conclusions that the indicated changes truly are likely to cause desired improvements. As such, to claim that a change would lead to some improvement is to make a causal claim, something that the oversight purpose does not require.

Corresponding to the three evaluation component tasks depicted in Figure 1 as contributing to program improvement, these causal claims often are supported by information about implementation, about outcomes, or about the underlying causal mechanisms responsible for program impact. In the first of these types, program improvement can be served when assessment reveals that the program has not been implemented as planned and there is reason to believe that the observed variances from the original plan are problematic. This last part is critical because sometimes variances from plans represent adaptive changes to fit local contexts. For example, hospital administrators might use observation to determine that a new patient intake procedure is not being implemented as planned. A subsequent patient satisfaction survey might confirm that the variations from the intended intake procedure are causing problems, and this would then provide guidance for making changes to improve the patient intake process.

With regard to using information about outcomes, a public school district might be confronted with disappointing results on a standardized mathematics test for high school

sophomores. School administrators and teachers might then discover that the low test scores in mathematics were from new students moving in to the district, that the long-term resident students are performing well. This information could lead to a more focused intervention for improvement than simply increasing time spent on mathematics for all students.

These examples involve incremental midcourse changes to ongoing programs based on at least some understanding of the dynamics involved. If the changes do not seem to resolve the identified concern, or seem to make things worse, then other changes can be attempted. A more involved approach to program improvement makes an even greater effort to understand the underlying causal factors, or mechanisms, that are responsible for current problems and might support desired improvements. In the education example, school administrators might study the characteristics of those not doing well on the mathematics test and might determine that one segment involves new students who feel unconnected to the school and are unmotivated, while another segment is speaking English as a second language and is performing poorly on other test subjects as well. The presumption of this concern with underlying mechanisms is that understanding why a problem exists might lead to more focused efforts at program improvement. Note that other efforts at understanding mechanisms can involve theory-based basic research (e.g., cognitive research studying the mechanisms underlying learning) and so would fall under the knowledge development purpose and would be sponsored at a higher level as discussed next.

Policymakers at a higher administrative level address a form of program improvement when reauthorization of a policy involves making changes to rectify identified problems with the original regulations or laws. With regard to studying underlying program mechanisms, it is common for federal agencies to fund basic research to uncover program-related mechanisms. Consumers and citizens become involved in program improvement efforts when advisory boards with their representation take on this purpose, in addition to the program oversight purpose described above.

Strategic Resource Allocation. The third purpose is concerned with making major allocation decisions regarding programs. This often involves selecting between and among existing alternatives, as opposed to making changes in an ongoing program. This is the domain of summative evaluation, and it is associated with the arrow in Figure 2 that involves using

feedback from the evaluation-based understanding to inform decisions about the resources allocated to programs. This is the focus of those who would use performance measurement for strategic management decisions. One exemplar for this purpose is performance measurement, or another form of evaluation, is used to determine whether resources allocated to a program, such as a teen pregnancy prevention program, should be allocated instead to another program, such as an outreach program for infant vaccination. If the resulting judgment is that the program is effective and cost-effective, it might be continued, expanded, or institutionalized with broad dissemination. If not, the program might be curtailed or cancelled (though experience suggests that programs are rarely cancelled due to poor evaluation results). Note that this purpose requires three things, first, good information about the nature of the unmet needs in a community or other area, second, confidence about the judged effectiveness of alternative program, and, third, some understanding of how the public values the alternative program outcomes (e.g., teen pregnancy prevention versus healthier babies). Performance measurement has methodologies to address the first two of these but leaves it to administrators and other stakeholders to make the value judgments about alternative outcomes.

Another exemplar of resource allocation involves consumers choosing to participate in a program or not, something that is analogous to consumers choosing to purchase one product rather than another. As when choosing to buy a particular car, choosing to enroll one's children in an afterschool program or a designated magnet school is for the most part an all-or-nothing decision with little opportunity to provide input for program improvement. For such decisions, consumers need to have a good sense of the consequences that might be expected from enrollment in a particular program if they are to reach meaningful judgments about the value such program have for them. Similarly, citizens who are presented with referenda on policy changes should have some information to support their judgments.

Program managers do get involved in allocation decisions but at a more tactical, or local, level. For example, school administrators and teachers evaluate and select textbooks, and the manager of a health clinic can evaluate and select suppliers of the medical products used. These decisions, however, tend not to be at the same strategic level as those called for in large scale program funding decisions or policy reforms, such as welfare reform.

Choosing among Evaluation Purposes

The emphasis on choosing a primary, and perhaps secondary, purpose for an evaluation reflects the limits confronted by practicing evaluators—it is rare to have the resources necessary to pursue all three purposes in a single evaluation and so some choices are necessary. How are we to make these choices? The short answer is that we want to select evaluations purposes that are appropriate in given contexts. The longer answer requires a framework that specifies which aspects of the context are key in leading to one purpose over another. While not a simple matter, this task of identifying factors that condition the appropriateness of standard evaluation practices is at the heart of most recent evaluation theories (see Mark et al., 2000; Shadish, Cook, & Campbell, 2002; and Shadish, Cook, & Leviton, 1992). There currently are no firm principles on how to resolve this tension, but the following three issues warrant consideration.

Choosing according to identified stakeholders. As indicated in Table 1, different stakeholder groups tend to be most interested in particular questions. So, one way to choose an evaluation purpose would be to focus on the needs of a particular group of stakeholders, such as those most likely to use, to act on, the findings of the evaluation. This is the guiding insight of Patton’s utilization focused evaluation, wherein the focus on “intended uses by intended users” is presented as a reasonably sufficient maxim for guiding decisions about evaluation purpose. Thus, if a funding agency is identified as the primary stakeholder for an evaluation of a new program, discussions might lead to program oversight being identified as the primary purpose, whereas an evaluation commissioned by program managers might be more focused on recommending ways to improve the program. The point to emphasize is that the appropriate purposes do not emerge from the classification of stakeholder groups but from identifying actual people and engaging in enough dialogue to understand their specific information needs (Patton, 1997). If this dialogue reveals that the primary stakeholders cannot agree on what the evaluation should attempt to achieve, any resulting evaluation is likely to lack focus and its results likely to yield little use (Rossi et al., 2004).

While maintaining a tight focus on the identified primary stakeholder offers many virtues, there is some danger of neglecting purposes that are particularly important for other

stakeholders. Indeed, maximizing the usefulness of an evaluation for one group of identified users might make the evaluation results relatively useless for other potential users. This is a specific point in a larger critique offered by House (1991) as a warning to avoid six ethical fallacies regarding the value foundations of evaluation. The first that he addresses, clientism, is particularly relevant here in that it refers to the dysfunction of focusing an evaluation on only what is of interest to the identified evaluation client (generally, the people who are paying for the evaluation). Related, contractualism involves restricting the evaluation purpose to what is specified in the contract for work; managerialism is restricting the purpose of the evaluation to the needs of program managers; and pluralism/elitism involves giving priority in addressing the needs of one or more powerful groups.

House calls these value stances ethical fallacies to highlight what he sees as the mistake of privileging automatically the values of one group over another. On the other hand, a fifth fallacy identified by House is relativism, the mistake of accepting the values of all stakeholder groups as necessarily equally valid. The general point of his critique is that evaluators cannot simply abdicate responsibility for the choice of values underlying an evaluation. Working with stakeholders to derive primary, and perhaps secondary, purposes for evaluations is certainly important, but we need to consider as well the values and needs of stakeholders who are not part of that dialogue.

To complete consideration of House's critique, his sixth fallacy is methodologicalism, or giving privilege to a particular evaluation methodology, regardless of the needs of the context. As I hope is clear, avoiding this fallacy by instead selecting methods appropriate for a given context is the central goal of this paper.

Choosing according to project stage of development. In addition to the issues that follow from assigning priority to one evaluation user over another, most would identify the program stage of development as key in fitting evaluations to the needs of stakeholders (see Tharp, 1981). The rationale here is that one could construct a program theory, or logic model, to represent the sequential logical steps by which program-related activities lead to intermediate and then valued outcomes, and then, hopefully, to a better society. Viewing these steps along a chronological timeline, in the early stages the primary evaluation purpose is likely to be

oversight, with an emphasis on ensuring that program implementation is as intended. Once the program activities are in operation, the program improvement purpose is likely to dominate, with quick feedback leading to changes in program activities. Only after the program appears to be performing at something near its full potential is it useful to for allocation questions involving program selection to be addressed, with the consequences of modifying program resources to expand, maintain, or downscale the program. As Rossi et al. (2004, p. 39) point out, attempting to make selection decisions too soon typically places unrealistic expectations on what a developing program can be expected to achieve.

In sum, we would not expect that the questions which are foremost when a pilot project is being developed will be the same as those that are central when considering whether to disseminate practices or policies from an apparently successful pilot project. This developmental view, however, interacts with the funding context of a program. A program with a guaranteed, stable source of funding is in a better position to follow this development sequence of evaluation than is a program for which decisions on possible termination are made yearly from the first year of program implementation (Tharp, 1981).

Choosing according to available resources. An additional contextual factor raised by Rossi et al. (2004) is the constraint that comes from considering the resources available for conducting an evaluation. One resource relevant for any evaluation, of course, is the funding available to support the evaluation. Limited funding generally requires a narrower focus for the evaluation and less emphasis on the elaborate designs that tend to be used for strategic allocation decisions. As Rossi et al. point out, however, other necessary resources include personnel and time. If the evaluation expertise is not available to monitor outcomes, there may be limits to the ability to promote program improvement. Equally important is the availability of program management personnel able to support data collection, access to clients, staff, and archival material, and implementation of actions suggested by the evaluation. Time as a constraint is relevant in that it takes time to develop a monitoring system for program improvement and even more time to set up a rigorous comparison design evaluation that would support resource allocation decisions.

B. Tasks of Evaluation

The main justification for differentiating evaluation purposes is that giving priority to one purpose over another has implications for what you are to do as the evaluator. Thus, the parallel claim is that performance measurement needs to be conducted differently when employed in support of different purposes. In this section we focus on the tasks addressed by three standard evaluation activities—implementation evaluation, outcome monitoring, and impact evaluation—and connect them to the contexts and corresponding purposes for which they are most useful.

Distinguishing the Evaluation Tasks of Performance Measurement

There is no one way to organize the various tasks undertaken by evaluators to address the diverse questions posed, but some organization is useful. Some tasks are required in all evaluations, others are appropriate only for certain purposes. Rossi et al. (2004) distinguish five general types, or “forms” in their terms, of evaluation tasks: Needs assessment, Assessment of program theory, Process evaluation, Impact assessment, which includes outcome monitoring, and Efficiency assessment, which addresses the worth or value of a program. Weiss, on the other hand, delineates the domain of evaluation in terms of the 14 questions (1998, p. 278) that evaluators often address. Table 2 presents this list but categorizes the questions into five broad groupings that helps us see the relationship with the Rossi et al. framework. If we distinguish outcome monitoring from impact assessment and further recognize that the approach to needs assessment that uses performance measurement is a form of outcome monitoring, we can delineate three evaluation tasks that are particularly relevant to performance measurement, as depicted in Figure 1: implementation evaluation, outcome monitoring, and impact assessment.

Table 2: Weiss’s Evaluation Questions, Organized by Evaluation Tasks
<p>A. Implementation Evaluation</p> <p><i>“What went on in the program over time?”</i></p> <p><i>“How closely did the program follow its original plan?”</i></p>
<p>B. Outcome Monitoring</p> <p><i>“Did recipients improve?”</i></p> <p><i>“Did recipients do better than nonrecipients?”</i></p>
<p>C. Impact Assessment</p> <p>Aggregate Program Impacts</p> <p><i>“Is the observed change due to the program?”</i></p> <p>Disaggregated Impact Assessment</p> <p><i>“What characteristics are associated with success?”</i></p> <p><i>“What combinations of actors, services and conditions are associated with success and failure?”</i></p> <p>Assessment of Causal Mechanisms</p> <p><i>“Through what processes did change take place over time?”</i></p>
<p>D. Assessment of Worth</p> <p><i>“What was the worth of the relative improvement of recipients?”</i></p>
<p>E. Critical Review</p> <p><i>“What unexpected events and outcomes were observed?”</i></p> <p><i>“What are the limits to the findings?”</i></p> <p><i>“What are the implications of these findings? What do they mean in practical terms?”</i></p> <p><i>“What recommendations do the findings imply for modifications in program and policy?”</i></p> <p><i>“What new policies and programmatic efforts to solve social problems do the findings support?”</i></p>

Implementation Evaluation. This form of evaluation involves documenting the activities of the program and the resources that were used to support those activities. Weiss’s questions ask about the program activities and about the consistency of program activities with those described

in the original program plan. Rossi et al. present a list of questions for implementation, or process, evaluation (2004, pp. 171-172) that require both the description emphasized in Weiss's questions (who is receiving services?; what services are they receiving?) as well as broader range of judgments (is the program well organized?; are the resources used effectively and efficiently?).

The descriptive elements of implementation evaluation are straightforward, with the key challenge being to measure the right things correctly. Making judgments in implementation evaluation is usually of one or both of two types, either of comparing activities to those authorized or planned or of judging the activities in the context of what are understood as "best practices." This supports the oversight purpose by calling attention to what are considered problems that require attention. The program improvement purpose can also be served with this information as a first step in reaching conclusions about activities that would lead to improvement.

Outcome Monitoring. Outcome monitoring has been controversial in the evaluation field, in part because of a tendency to neglect the distinction between the methodology being used and the purposes for which the methodology is employed. One use of outcome monitoring is to support oversight by those responsible for programs, including both administrators and the citizenry. In typical practice, outcome monitoring is used to make simple comparisons over time to track changes. For example, having identified adolescent pregnancy as a problem, a state agency might track the number of births to adolescent mothers per 1000 adolescent girls to see if the rate is increasing, remaining stable, or decreasing over time. Outcome monitoring can also be used as part of needs assessment where indicators, such as the number of premature births per 1000 births, are used to generate a profile of relative and absolute needs. Boruch, De Moya, and Snyder (2002, p. 53) discuss survey approaches to outcome monitoring that can be used to support needs assessments in the areas of delinquency, illiteracy and crime.

If used in this strictly descriptive manner, outcome monitoring involves no claim about the causal impact of a program or of a change in a program. Rather, consistent with the program oversight purpose, this descriptive information could help direct attention to identified problems that are getting worse. Much of the controversy about outcome monitoring comes when it is

used to go beyond simple description to support claims about program improvement and resource allocation. For example, it is problematic to use outcome monitoring to support the conclusion that a policy change (e.g., a policy change for Social Security Disability Insurance) is responsible for an observed change in an outcome (e.g., an apparent decrease in the employment rate of persons with disabilities; see Silverstein, Julnes, & Nolan, 2005). To support such causal conclusions, more is demanded of the evaluation methodology.

For outcome monitoring to be useful in service of program improvement, we need to have some confidence that identified changes would indeed improve program functioning and outcomes. One approach is to make use of more sophisticated comparisons. For example, outcome monitoring can represent relative trends wherein the changes over time for the one group, such as the adolescent pregnancy rate for one state, follow a different pattern than those for other groups, such as a dozen other states selected to be comparisons for different dimensions of similarity. Much of the No Child Left Behind initiative in education depends on the meaningfulness of these types of relative trends across schools in the U.S. Continuing this elaboration, if standardized test scores are available for many years, covering a broad range of content areas, and across all school districts in a state, or even across the country, differential changes in one content area and in only certain districts can become easier to interpret.

When the primary evaluation purpose is resource allocation, outcome monitoring will not be adequate on its own. The reason is that for selection between alternatives we need to have some understanding of the sizes of the various impacts of the alternative programs. This is the focus of impact evaluation.

Impact Evaluation. Assessing the impact of a program is a central and yet difficult task in an evaluation. As Weiss (1998) notes, a central question is—“Is the observed change due to the program?”. Some of the analyses conducted for impact evaluation are similar to those used for outcome monitoring, such as comparing outcomes for two or more groups. As such, both tasks depend on accurate description. The main difference is that the goal of design and analyses in impact evaluation is to rule out alternative explanations for impact evaluation. This goal places a consequent greater burden on impact evaluation to go beyond description and in addition to support causal conclusions. A related concern is estimating the size of the program impact, a

task that requires even greater confidence in ruling out alternative causal influences.

How this ruling out of alternative explanations is done through design and analysis is, of course, central to evaluation. The point here is that performance measurement needs to address this ruling out of alternatives but need not be judged inferior to full-scale impact evaluations in this regard. That is, it is important to distinguish methods, tasks, and purposes. Using performance measurement to represent outcomes for oversight is different than using it to support the causal conclusions necessary to claim that a program is effective or that certain changes will “cause” a program to be more effective. For performance measurement to contribute to the impact evaluation task it must be conducted strategically. If, for example, multiple related outcomes are monitored but only the outcomes targeted by a program show improvement (i.e., the other outcomes showed no improvements), we would feel more confident that the improvements were program-related. Similarly, if the program were implemented sequentially in multiple districts in a city with a sufficient lag time between each district implementation and the monitored outcomes showed improvement only after implementation (e.g., three months after implementation in each district), we would again have more confidence that the observed changes in outcomes were program-related. There are many other circumstances where performance measurement can support confident causal conclusions, but the key is being strategic, just as one needs to be in conducting any impact evaluation. Which alternative explanations are most important to rule out, will generally depend on the context in which the performance measurement system is being implemented.

Taking this point further, impact evaluation often addresses, as indicated in Table 1, several related questions. First, as noted by Weiss, evaluators are often concerned with moderated relationships wherein the size of the impact of a program is dependent on other factors. Quantitative analyses of moderated relationships typically make use of interaction terms in multiple regression and ANOVA. A second ancillary question addresses what are believed to be the underlying mechanisms responsible for the observed effects. Gathering evidence about those mechanisms, or processes, allows one to reassess the program theory that guided the program and to suggest modifications to the program theory. In traditions such as regression analysis and structural equation modeling, the study of mechanisms is the study of mediated relationships, or

the causal linkages, that relate, in whatever detail is desired, the program activities to the targeted outcomes. As with the initial point made about impact evaluation, performance measurement can help answer these ancillary questions, but only if it is designed and implemented in a strategic manner that supports disaggregating outcomes by relevant groups or allows some insight into the causal linkages that are responsible for observed outcome patterns. Both of these strategic intents require being sensitive to the demands of particular contexts.

C. Summary

Performance measurement, like any evaluative approach, can serve different purposes. To do so well requires first being aware of the alternative purposes and of which is most central in given circumstances. Working from an identified primary purpose, performance measurement can contribute to the evaluative tasks that are required for that purpose, again by being strategic in designing and implementing the performance measurement system to meet the needs of a given context. This paper supports this strategic approach by offering a systematic framework that connects performance measurement to context-sensitive developments in other areas with the field of evaluation.

References

Berman, E. M. and Wang, X. (2000). Performance measurement in U.S. counties: Capacity for reform. *Public Administration Review* 60 (5), 409-420

Boruch, R., de Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (eds.), *Evidence matters: Randomized trials in education research* (pp. 50-79). Washington, DC: Brookings.

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice, *Administrative Science Quarterly*, 17, 1-25.

de Lancer Julnes, P. (2006). Engaging citizens in governance-for-results: Opportunities and challenges. In M. Holzer (ed.), *Citizen-driven performance*. Seoul Development Institute: Seoul, Korea.

de Lancer Julnes, P. (in press) , Performance Measurement: An Effective tool for government accountability? The debate goes on. *Evaluation- The International Journal of Theory, Research and Practice*.

de Lancer Julnes, P. & 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.

de Lancer Julnes, P. & Mixcoatl, G. (in press). Governors as agents of change: A comparative study of performance measurement initiatives in Utah and Campeche. *Public Performance and Management Review*.

Dusenbury, P., Blaine, L., & Vinson, E. (2000). *States, Citizens, and Local Performance Management*. Urban Institute: Washington, DC.

Gramlich, E. M. (1990). *Benefit-cost analysis for government programs* (2nd ed.) New York: McGraw-Hill.

Halachmi, A. (2002). Who gets what when and how: performance measures for accountability? for improved performance? *International Review of Public Administration*, 7(1), 1-11.

House, E. R. (1991). Realism in research. *Educational Researcher*, 20, 2-9.

Julnes, G. & Foster, E. M. (2001). "Crafting evaluation in support of welfare reform." In G. Julnes & E. M. Foster (eds.), *Outcomes of welfare reform for families who leave TANF*. New Directions for Evaluation, no. 91, pp. 3-8. San Francisco: Jossey-Bass.

Julnes, G., & Mark, M. (1998). Evaluation as sensemaking: Knowledge construction in a realist world. In G. Henry, G. Julnes, & M. Mark (Eds.) *Realist evaluation: An emerging theory in support of practice*. New Directions for Evaluation, no. 78 (pp. 33-52). San Francisco: Jossey-Bass.

Kelly, J. M. (2002). "If You Only Knew How Well We Are Performing, You'd be Highly Satisfied With The Quality Of Our Service." *National Civic Review* 91 (3), p. 283-292.

Kelly, J. M. & Swindell, D. (2002). "Service quality variation across urban space: First steps toward a model of citizen satisfaction". *Journal of Urban Affairs* 24 (3), 271-288.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving public and nonprofit policies and programs*. San Francisco: Jossey-Bass.

Newcomer, Kathryn, ed. 1997. *Using performance measurement to improve public and nonprofit programs*. San Francisco: Jossey-Bass.

Okun, A. M. (1975). *Equality and efficiency: The big tradeoff*. Washington, DC: Brookings Institution.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*, 7th edition. Thousand Oaks, CA: Sage.

Scriven, M. S. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. S. Scriven (Eds.), *Perspectives of curriculum evaluation* (AERA Monograph Series on Curriculum Evaluation, No. 1, pp. 39–83). Skokie, IL: Rand McNally.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Tharp, R. G. (1981). The metamethodology of research and development. *Educational Perspectives*, 20, 42-48.

Silverstein, R., Julnes, G., & Nolan, R. (2005) What policymakers need and must demand from research regarding the employment rate of persons with disabilities. *Behavioral Sciences and the Law*, 23, 1-50.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies*. Upper Saddle River, NJ: Prentice Hall.

Wholey, J. S. (1997). Trends in performance measurement: Challenges for evaluators. In E. Chelmsky & W. R. Shadish (eds.), *Evaluation for the 21st century: A handbook* (pp. 124–133). Thousand Oaks, CA: Sage.

Wholey, J. S. (1983). *Evaluation and Effective Public Management*. Boston: Little, Brown.

Wholey, J. S. (1979). *Evaluation: Promise and performance*. Washington, DC: Urban Institute.