Performance Measurement Challenges and Strategies
(June 18, 2003)

I. Introduction

This document provides practical strategies for addressing common performance measurement challenges. It grew out of the workshop on performance measurement organized by the Office of Management and Budget (OMB) and the Council for Excellence in Government which was held on April 22, 2003.

The document is meant to complement the Program Assessment Rating Tool (PART) guidance document (www.omb.gov/PART), which also addresses performance measurement. Following this introduction, Section II discusses basic performance measurement definitions and concepts. Section III then discusses six common performance measurement problems that were the subject of break-out sections at the workshop.

Many of the performance measurement issues that Federal program managers face are extremely difficult, and this document offers no easy solutions. Rather, this paper suggests some potentially useful strategies for addressing these issues. Suggestions on additional challenges, strategies, and examples are welcome, so that this document can evolve. Suggestions may be sent to performance@omb.eop.gov or to any member of OMB's Performance Evaluation Team.

Performance measurement indicates what a program is accomplishing and whether results are being achieved. It helps managers by providing them information on how resources and efforts should be allocated to ensure effectiveness. It keeps program partners focused on the key goals of a program. And, it supports development and justification of budget proposals by indicating how taxpayers and others benefit.

However, information provided by performance measurement is just part of the information that managers and policy officials need to make decisions. Performance measurement must often be coupled with evaluation data to increase our understanding of why results occur and what value a program adds. Performance measurement cannot replace data on program costs, political judgments about priorities, creativity about solutions, or common sense. A major purpose of performance measurement is to raise fundamental questions; the measures seldom, by themselves, provide definitive answers.

Because performance measurement keeps a focus on results, it has been a central aspect both of the Government Results and Performance Act (GPRA) and of the PART. One goal of the PART is to try to ensure that the most relevant performance information is readily accessible to policy makers.

The PART seeks to answer whether a program is demonstrating value to the taxpayer. In doing so, the PART sets a standard for performance information that is high

but also basic and compelling. Ideally, it seeks to demonstrate that a program 1) has a track record of results and 2) warrants continued or additional resources.

We are far from having the data and ability to do such analysis on the full range of Federal programs. But, the identification of adequate performance measures is a necessary step in integrating performance information and budget decisions.

II. Key Definitions and Concepts

1. Definitions used in the PART

Strategic goals are statements of purpose or mission that agencies may include in a strategic plan. Strategic goals might not be easily measurable. For example, a strategic goal for a weather program might be protecting life and property, and promoting commerce and the quality of life, through accurate forecasts. To the greatest extent reasonable, the PART encourages agencies to use their strategic goals to develop specific, operational performance goals.

Performance goals are the target levels of performance expressed as a measurable objective, against which actual achievement can be compared. Performance goals can be stated as either **outcomes** or **outputs**, but to be complete they should incorporate targets and timeframes into a performance measure.

- **Performance measures** are the indicators or metrics that are used to gauge program performance. Performance measures can be either outcome or output measures. Using again the example of a weather program, a measure might be average advance warning time for tornadoes. Performance measures correspond with questions 2.1 and 2.3 in the PART.

- **Targets** are the quantifiable or otherwise measurable characteristics that tell how well a program must accomplish a performance measure. The target for tornado warning time, for example, might be an average of 20 minutes by the year 2008. Targets correspond with questions 2.2 and 2.4 in the PART.

In summary, together with the performance measure, the targets and timeframes establish a performance goal. For the weather program example, the performance goal would be an average tornado warning time of 20 minutes by 2008.

The PART requires two types of performance goals:

- **long-term performance goals** address performance that is generally several years or more in the future. There are two basic types of long-term goals: 1) an annual performance goal in the future, (e.g., tornado warning times in 2008, or unit costs of an activity in 2010); and 2) the cumulative effect of annual activities (e.g., development of an AIDS vaccine by 2010). Long-term program goals are

required under both GPRA (termed "general goals") and the PART (questions 2.1 and 2.2).

- **annual performance goals** should be stated in yearly increments (questions 2.3 and 2.4). For the weather program example, an annual performance goal might include the same performance measure (advance warning time), but a less ambitious target (e.g., 15 minutes average warning time in 2005) due to less widespread use of advanced technologies.

2. Outcomes, Outputs, and Inputs

Outcomes describe the intended result or consequence that will occur from carrying out a program or activity. Outcomes are of direct importance to beneficiaries and the public generally. While performance measures should distinguish between outcomes and outputs, there should be a logical connection between them, with outputs supporting outcomes in a logical fashion. The PART strongly encourages the use of outcomes because they are much more meaningful to the public than outputs, which tend to be more process-oriented or means to an end. Outcomes may relate to society as a whole or to the specific beneficiaries of programs, depending on the size and reach of the program.

Example (see 2004 PART for Maternal and Child Health Block Grants (MCHBG)):
- *Long-term measure: National rate of maternal deaths per 100,000 live births in 2008.*
- *Annual measure: National rate of illnesses and complications due to pregnancy per 100 deliveries in 2004.*

It is sometimes not possible to measure outcomes annually. In these cases, it is likely that output goals will be used for annual measurement.

Example: An outcome goal for a space program might be to determine whether there is life on Mars by 2011; annual goals, however, might relate to accomplishing steps toward developing the exploration vehicle and systems.

Outputs are the goods and services produced by a program or organization and provided to the public or others. They include a description of the characteristics and attributes (e.g., timeliness) established as standards.

Example (see 2004 MCHBG PART):
- *Number of Medicaid-eligible children who receive MCHBG services.*

Managers are more likely to manage against outputs rather than outcomes. This is because output data is collected and reported more frequently, and outputs more typically correspond to activities and functions being directly controlled, as opposed to focusing on results. Nevertheless, outputs should help track a program's progress toward reaching its outcomes.

Outputs can include process measures (e.g., paper flow, adjudication), attribute measures (e.g., timeliness, accuracy, customer satisfaction), and measures of efficiency. They may be measured either as the total quantity of a good or service produced, or may be limited to those goods or services having certain attributes (e.g., number of timely and accurate benefit payments). Typically, outputs are measured at least annually.

Inputs are resources, often measured in dollars, used to produce outputs and outcomes. Performance measures may include consideration of inputs, particularly in the context of cost-efficiency or unit costs. Programs are encouraged to consider the most meaningful level of such input measures. For example, cost-efficiency measures based on outputs per dollar will typically be more useful than measures of output per unit of personnel (such as Full Time Equivalents). Similarly, social costs may be more meaningful than Federal budget costs when evaluating effectiveness of regulatory programs. Inputs from State and local partners may be relevant in assessing the effectiveness of some programs matched by Federal assistance.

3. Characteristics of good performance goals

The key to assessing program effectiveness is measuring the right things. Performance measures should capture the most important aspects of a program's mission and priorities. Appropriate performance goals should: 1) include both performance measures and targets; 2) focus on outcomes, but use outputs when necessary; and 3) include both annual and long-term measures and targets.

Characteristics of good performance goals include:

- Quality over quantity. Performance goals should be relevant to the core mission of the program and to the result the program is intended to achieve. This generally argues for quality over quantity, with a focus on a few good measures. However, programs should not feel compelled to collapse complex activities to a single measure, particularly if that measure is a proxy for the true objective.

- Importance to budget decisions. Performance goals included in the PART should provide information that helps make budget decisions. Agencies can maintain additional performance goals to improve the management of the program, but they do not need to be included in the PART.

- Public clarity. Performance goals should be understandable to the users of what is being measured. Publicize (internally and externally) what you are measuring. This also helps program partners understand what is expected from the program.

- Feasibility. Performance goals should be feasible, but not the path of least resistance. Choose performance goals based on the relevancy of the outcomes and not for other reasons -- not because you have good data on a less relevant

measure, for example. If necessary, terminate less useful data collections to help fund more useful ones.

- Collaboration. Agencies and their partners (e.g., States, contractors) need to work together and not worry about "turf" – the outcome is what is important.

4. Getting Started

Defining the right performance measures can sometimes be like talking to a four-year-old child – whatever you say, the response is always "Why? Why? Why?" Similarly, getting to a good measure can often grow out of asking why a certain activity, input, or output is important and what it is really trying to achieve that matters to the public. This kind of drilling down to get to the right outcome measure might look like this for a job training program:

Example: Possible Measures for Job Training Programs

o Dollars appropriated to the program o Number and size of grants *Why do these matter? What do they buy?*	Inputs: Funding (Federal and perhaps State and local)
o Number of classes attended by program participants o Number of people trained *Why do these matter? What result do they produce?*	Outputs: Products (e.g., classes taught, service delivered, participants serviced)
o Number of people with useful skills o Number of people who get a job after leaving the program *Why do these matter? Is this the result the public is seeking?*	Intermediate outcomes: (e.g., new knowledge, increased skills, changed behavior)
o Number of program participants who remain employed for a specified time and increase their earnings	Program outcome
o Number of people who are self-sufficient	Societal outcome

Considering the scope of a program is also key to identifying proper performance measures. For example, output goals were used in the 2004 PART for the U.S. Fish and Wildlife Service (USFWS) National Fish Hatchery System (NFHS) because of the difficulties in attributing success in achieving species conservation goals – a higher level outcome – based solely on propagation of hatchery fish. Success at the outcome goal of species conservation would be better assessed by considering a broader scope, such as the USFWS Fisheries Program, which includes both the hatchery (NFHS) and habitat improvement aspects of species conservation. In addition, while external factors such as

other stakeholders' actions and drought also affect species conservation, the Fisheries Program can take these into account as it develops its goals and carries out its activities.

III. Common Performance Measurement Issues

Based on the April 22nd workshop and follow-up discussions, this portion of the document outlines six common performance measurement issues and offers possible strategies for addressing them. The issues address programs that: 1) have outcomes that are extremely difficult to measure; 2) are among many contributors to a desired outcome; 3) have results that will not be achieved for many years; 4) relate to deterrence or prevention of specific behaviors; 5) have multiple purposes and funding that can be used for a range of activities; and 6) are administrative or process oriented.

Whenever possible, the document provides examples of performance goals and 2004 PARTs that effectively address the problem at hand. All PART summaries and the completed PART worksheets can be found at http://www.whitehouse.gov/omb/budget/fy2004/pma.html.

1. The program's outcomes are extremely difficult to measure

Some programs' outcomes are inherently difficult to measure. For example, programs designed to address foreign policy objectives might fall into this category. By focusing on why a program is important and what makes it difficult to measure, the scope of the problem can sometimes be more specifically defined. Going through this process may also identify the root of the 'difficult to measure' problem as one of the other core problems identified in this document.

Performance measure challenges can often be traced back to fundamental questions about the program, which when reexamined may yield insights into better ways to assess effectiveness. As mentioned earlier, one way to reexamine those issues is to relentlessly ask "why?"

- Why it is important that the program receive funding?
- Why are program operations important?
- Why does the program do what it does?
- If the program were fabulously successful, what problem would it solve?
- How would you know?

This line of questioning should help clarify the program's true purpose and what its desired outcome is, which should help determine what really needs to be measured. For example, a program's purpose may be to support an international coalition. In trying to define a performance measure, it might be helpful to ask "Why is the success of that coalition important and what role does the program play in achieving that goal?"

It also can be helpful to identify what core issues make measurement difficult. For example:

- The program purpose is not clear.
- The beneficiary or customer is not defined. Consider who are the direct and indirect beneficiaries. Who are the long- and short-term beneficiaries? If the government does not do this, who would pay for it?
- Stakeholders have a different view of the program than program managers. How would stakeholders be affected if the program did not exist? Are there performance measures for stakeholders that shed light on the program's effectiveness?
- Some programs are difficult to measure because data is not available. To help address this situation, ask the following questions: Why is data unavailable? What data is available? Can we fund the cost to find data? If data is not available, are there proxy measures that will indirectly measure the program's outcomes? Do stakeholders have data that they generate to track the program?
- If quantitative data is unavailable and inappropriate, consider using qualitative data, such as assembling a panel of experts on the topic. For example, in assessing the quality of public defenders' services, a survey of judges may be useful, and could complement output measures such as cost per case.

2. The program is one of many contributors to the desired outcome

Often several Federal programs, programs from various levels of government (Federal, State, local), private-sector or non-profit activities, or even foreign countries all contribute to achieving the same goal. The contribution of any one Federal program may be relatively small or large. Examples of programs with these characteristics include international peacekeeping (PKO 2004 PART), special education pre-school grants (IDEA Preschool 2004 PART), highways (FHWA Highways 2004 PART), Vocational Education (2004 PART), and many education, labor, and housing formula grant programs.

One approach to this situation is to develop broad, yet measurable, outcome goals for the collection of programs, while also having program-specific performance goals. For a collection of programs housed primarily in one Federal agency, a broad outcome measure may be one of the goals in an agency strategic plan (e.g., increasing the home ownership rate). The broad outcome goal can often be tracked using national data that is already being collected, while the program-specific goals may require more targeted data collection. Both the broad outcome goal and the program-specific goals could be addressed in the PART.

Example: Several Federal education programs, totaling nearly $14 billion, contribute to helping children learn to read. One of those programs, Reading First State Grants, provides about $1 billion to help implement proven literacy reforms in schools with low reading scores.

- *Common outcome goal: Percentage of children in high-poverty schools reading proficiently by the end of third grade.*

- *Reading First goal: Percentage of at-risk third graders receiving Reading First services who can read at or above grade level.*

It is important to "right size" the measure to suit the program. Sometimes a program is such a significant contributor, or leverages so many dollars, that an appropriate goal is a societal outcome. Other times it is more appropriate to write measures specific to program beneficiaries. There is no rule of thumb on where that threshold is. We suggest only that programs of similar size, or with a similar percentage contribution to the desired outcome, approach this issue similarly.

Example: Several Federal programs provide student aid so that low and moderate income students can afford to attend college. Of these, only the Pell Grant program and the loan programs contribute a large enough share of student aid to merit a societal outcome. The Pell Grant program provides grants to nearly one-third of all college students, while about half of all students receive loans from or backed by the Federal government. In contrast, the College Work Study program reaches only about 6% of college students, and so the measures relate to the program participants only:

- *Federal Pell Grant long-term measure (see 2004 PART): College enrollment gap between low-income and high-income high school graduates.*

- *College Work Study long-term measure: Rate of College Work Study students who complete their post-secondary education program.*

Sometimes programs are designed to work together toward a common goal, but each provides a different piece of the service or activity. In other cases, programs are designed to merge funds and support the same activities as well as goals; this is particularly true when Federal, State, and local dollars all contribute to reaching a common goal.

When programs fund different activities and do not co-mingle funds, programs should be able to develop activity-specific performance goals that support the broader outcome. It is likely, however, that these will be output goals and the challenge will be agreeing on how each of the separate activities contributes to the outcome.

When programs co-mingle funds in support of a goal, it is extremely difficult to assess the marginal impact of the program dollar since all funding supports similar activities. Programs may seek to claim responsibility for the entire outcome and output, despite having a shared, and sometimes small, role in the overall activity. However, we should seek to evaluate whether such claims are realistic. It may be useful in such situations to consider measures such as unit costs in terms of output per Federal dollar spent as well as the output per combined Federal, State and local dollars spent.

There are three basic sets of questions that one would aim to answer with performance information:

- First, is the overall effort working? Are there outcome measures for the overall effort/program? Are there evaluations?
- Second, is the Federal contribution making a difference? Because withholding funding as an experiment is not a viable option, analysts should consider whether there are other ways of seeing what would happen in the absence of Federal funding. Can one compare current funding to an earlier time when there was no Federal funding? Are there regions of the country where there is no Federal funding?
- Third, how is funding effort shared between Federal and non-Federal partners? How does the distribution of funding effort compare to measures of need or the distribution of benefits?

3. Results will not be achieved for many years

In some cases, the outcome of a program may not be realized for many years. In some cases, this can be addressed by identifying meaningful output-oriented milestones that lead to achieving the long-term outcome goal. Many research and development (R&D) programs, such as Hydrogen Technology (2004 PART) and Mars Exploration (2004 PART), fall into this category.

To address this issue, a program should define the specific short- and medium-term steps or milestones to accomplish the long-term outcome goal. These steps are likely to be output-oriented, prerequisite accomplishments on the path toward the outcome goal. A road map can identify these interim goals, suggest how they will be measured, and establish an evaluation schedule to assess their impact on the long-term goal. It is important that these steps are meaningful to the program, measurable, and linked to the outcome goal.

Example: *The purpose of NASA's Mars Exploration program is to explore Mars, focusing on the search for evidence of life. To that end, NASA defines spacecraft missions, which provide one level of measures to assess program effectiveness: mission success. Further, within each Mars mission, the program develops technologies; builds, launches, and operates robotic spacecraft; and performs research using the spacecraft instruments. While these steps take many years to complete, they provide many milestones against which a mission – and the program – can be monitored. Useful measures could include timeliness in achieving certain steps as well as percentage cost overruns.*

It may also be useful to track process-oriented measures, such as the extent to which programs make decisions based on competitive review. For example, research programs can have many uncertainties, including their expected outcomes. So, while research programs are encouraged to define measures that can track progress, not all will be able to. Such programs may rely, in part, on process measures, such as the extent to which the program uses merit-based competitive review in making awards.

To qualitatively address the research itself, some programs develop measures to reflect meaningful external validation of the quality and value of the program's research. To address the uncertainty of research outcomes, programs may also be able to demonstrate performance in terms of the broad portfolio of the efforts within the program. Expert independent evaluators might also help determine if the process of choosing appropriate long-term investments is fair, open and promises higher expected payoffs in exchange for higher levels of risk. Rotating evaluators periodically may help ensure independence and objectivity.

Another solution is estimation of future results using computer models or expert panels. EPA uses the former to estimate cancer cases avoided.

4. The program relates to deterrence or prevention of specific behaviors

Programs with a deterrence or prevention focus can be difficult to measure for a variety of reasons. Most importantly, deterrence measurement requires consideration of what would happen in the absence of the deterrence program. Also, it is often difficult to isolate the impact of the individual program on behavior that may be affected by multiple other factors.

Sample programs: Coast Guard drug interdiction (2004 PART), Department of Labor/Office of Federal Contract Compliance (2004 PART), Nuclear Regulatory Commission's Inspection and Performance Assessment Program.

If performance measures reflect a continuum from lower-level outputs to higher-level outcome measures related to the overall strategic goal, it is important for deterrence programs to choose measures that are far enough along the continuum that they tie to the ultimate strategic goal as well as to the program's activity. This will help ensure that the measures are both meaningful and genuinely affected by the program. Care should be taken, as some measures may create perverse incentives if they do not reach the correct balance between output and outcome (e.g., measures that focus on enforcement actions, as opposed to crime rates).

Example: A useful measure for the Coast Guard drug interdiction program could be the total volume of drugs entering the United States. This measure might be contrasted with drug seizure rates. High drug seizure rates might suggest that the Coast Guard interdiction strategies are effective. However, if the amount of drugs being sent rises significantly, and the number of seizures goes up to a lesser extent, the measure would still show that the Coast Guard program was effective, even though the volume of drugs getting through has increased substantially. In contrast, the total volume of drugs entering the U.S. is tied more closely to the overall strategic goal of reducing the flow of drugs into the country. On the downside, the Coast Guard has only partial control over the measure of volume entering the country.

Establishing deterrence targets. For some programs, deterring a majority of the negative outcome is appropriate. For other programs, most, if not all, of the negative

outcome must be avoided. In principle, the target for the program should reflect consideration of the maximization of net benefits (see, for example, OMB guidance on rulemaking under E.O. 12866). In any event, understanding the costs and benefits of compliance at the margins will help the program to determine the correct target level for compliance.

Example: For programs in which non-compliance is not life-threatening, and for which compliance is historically low, a legitimate long-term target may fall short of 100% compliance. In these cases, short-term targets that demonstrate forward progress toward the acceptable long-range goal may make sense.

Programs where failure is not an option. For programs where failure to prevent a negative outcome would be catastrophic (including programs to prevent terrorism or nuclear accidents), traditional outcome measurement might lead to an "all-or-nothing" goal. As long as the negative outcome is prevented, the program might be considered successful, regardless of the costs incurred in prevention or any close calls experienced that could have led to a catastrophic failure.

However, proxy measures can be used to determine how well the deterrence process is functioning. These proxy measures should be closely tied to the outcome, and the program should be able to demonstrate -- such as through the use of modeling -- how the proxies tie to the eventual outcome. Because failure to prevent a negative outcome is catastrophic, it may be necessary to have a number of proxy measures to help ensure that sufficient safeguards are in place. Failure in one of the proxy measures would not lead, in itself, to catastrophic failure of the program as a whole; however, failure in any one of the safeguards would be indicative of the risk of an overall failure.

Example: Outcome goals for the Nuclear Regulatory Commission of no nuclear reactor accidents, no deaths from acute radiation exposures from nuclear reactors, no exposure events at reactors, and no radiological sabotage are not necessarily sufficient to evaluate the program. There have been no occurrences of the above-mentioned events during the years 1999 to date. Therefore, annual goals used for the program include no more than one precursor event per year, no statistically significant adverse industry trends in safety performance, and no overexposures exceeding applicable regulatory limits. These proxy measures are useful to assess the ongoing effectiveness of this program.

5. The program has multiple purposes and funding can be used for a range of activities.

Some Federal programs are both large and diverse. They may be designed to address multiple objectives or support a broad range of activities or both. Block grant programs often have these characteristics, with the added feature of allowing grantees the flexibility to set priorities and make spending choices. Increased flexibility at the local level can limit efforts to set national goals and standards or create obstacles for ensuring accountability. In other cases, the program may focus on a limited set of activities which

in turn are used for multiple purposes by many distinct stakeholders. Establishing performance measures for these types of programs can be challenging.

Sample Programs: Block grants, such as the Community Development Block Grant program (CDBG), the Social Service Block Grant program (SSBG), and Temporary Assistance for Needy Families (TANF).

Establishing performance goals for block grant programs. Some block grant programs provide resources to non-Federal levels of government to focus on specific program areas, such as education, job training, or violence prevention. While the funds can often be used for a variety of activities, they are for a specific purpose. In these cases, national goals can be articulated that focus on outcomes to highlight for grantees the ultimate purpose of program funds. Targets for these measures may be set by surveying grantees to gauge the expected scale of their work or by looking at historical trend data. A system could be developed that uses performance measures and national standards to promote "joint" accountability for results. With this approach, after agreeing on an appropriate set of performance measures, program targets can be set at the local level and aggregated up to national targets.

Example: *CDBG is a large program with broad objectives. Seventy percent of CDBG funds are provided by formula to approximately 1,000 "entitlement" jurisdictions. The remaining 30 % of funds are allocated to States. The broad objectives of the program, coupled with local flexibility to determine community needs and relatively weak targeting criteria, have allowed grantees to use CDBG funds for a large range of activities. Through consultation with grantees and stakeholders, a core list of strategic objectives can be identified along with illustrative local/State performance measures. Each grantee could be asked to commit to specific strategic objectives and a set of procedures that will institutionalize a joint accountability partnership in each community. These procedures would be used to (1) establish and approve annual performance targets, (2) collect and verify performance data, and (3) determine when targets have been achieved. Accountability would require that results are publicized and assessed. A web-based system for reporting goals and accomplishments, for instance, could facilitate citizen review.*

6. The purpose of the program is administrative or process oriented.

Many programs in the government are administrative or process-oriented in nature which tends to present a number of problems when it comes to measuring performance. One issue is the appropriate balance between outputs and outcomes. Realistically, output measures may be useful for evaluating the efficiency of process oriented activities. In cases such as procurement of computer systems, for example, the spending may be better evaluated with other capital asset evaluation tools (such as business cases and Form 300s) than with the PART. However, for larger administrative efforts, consideration should still be given to ultimate outcomes. In some cases, it may make most sense to evaluate the administrative costs as part of the overall program, rather than as a separate activity. For example, a grant program may contain separate

accounts for the grants themselves and for administrative salaries and expenses, yet both accounts might be viewed as providing inputs into a single program .

Benchmarking with other agencies or the private sector, competitive sourcing, and the use of intermediate outcomes such as returns on investment are all approaches that can assist where data availability is an issue. For instance, GSA developed performance measures for its real property and vehicle acquisition programs based on private-sector costs.

As many administrative functions run across agencies, the development of common measures is also encouraged. For instance, the Inspector General community is working towards common measures to ensure consistency.

IV. Topics for Further Discussion

This is a living document that will be updated periodically to provide new ideas and to address new issues. To comment on this document or to raise additional issues for consideration, please send an e-mail to any member of OMB's Performance Evaluation Team or performance@omb.eop.gov.