NOAA Program Evaluation Guide

A publication of the NOAA Evaluation Committee

July 2, 2014

Table of Contents

INTRODUCTION

This guide outlines key steps for designing and implementing program evaluations. It provides practical guidance for managers responsible for the ongoing evaluation of NOAA programs and projects and can be used as a resource for staff with evaluation responsibilities. For project staff with evaluation experience, or who want to learn more, this guide provides enough basic information to conduct an evaluation. Staff will also find this guide helpful when working with an external evaluator to plan, design, and implement a successful evaluation.

This is not an exhaustive instructional guide for conducting program evaluations. It provides a framework for thinking about evaluation and outlines key steps for designing and implementing program evaluations. Sources of more detailed guidance on the technical aspects of evaluation can be found on the **NOAA Program Evaluation Network (PEN)** site.

What is a program?

There is no standard federal definition of "program." A program can be defined in various ways for budgeting and policy-making purposes. A program may be any activity, project, function, or policy that has an identifiable purpose or set of objectives. Programs can include:

- Direct service interventions
- Community mobilization efforts
- Research initiatives
- Observational systems
- Policy development activities
- Communication campaigns
- Infrastructure building projects
- Training and education services
- Administrative systems

What is program evaluation?

A program evaluation is a systematic study undertaken to assess how well a program is working and why. Program evaluations answer specific questions about program performance, and may assess a program's effectiveness, identify how to improve performance, or guide resource allocation. A program evaluation should be responsive to program and project needs, rather than an end in itself.

A program evaluation can assess an entire program or focus on a single initiative within a program. Program evaluations may also assess whether a program had unintended (perhaps undesirable) outcomes. Although evaluation of a federal program typically examines a broader range of activities than a single project, agencies may evaluate individual projects to identify effective practices or interventions.

Program evaluations are costly, infrequent events in the life of a program, and multiple factors should be considered in planning and implementing a program evaluation to ensure the effort makes a timely contribution to the program. Time and resource constraints often prevent design and implementation of the "best" evaluation: too many factors are in play when assessing large, complex programs to expect perfection. However, tried and true steps can be taken to improve the quality and utility of any program evaluation.

Program Evaluation and Performance Measurement

Sound program evaluations and performance measurements both contribute to the body of performance information that should guide the use of resources to improvement program performance.

Program evaluation is closely related to performance measurement, but important distinctions should be recognized. Performance measurement is the systematic, ongoing monitoring and reporting of program accomplishments, particularly progress toward performance goals. Performance measures or indicators may address program staffing and resources, the level of program activities conducted, the direct products or services delivered by a program, or the results of those products and services (**GAO, 2011**).

Unlike performance measurement, which spans the life of a program, program evaluations are discrete events designed to gather detailed information on a specific issue or issues. Program evaluations require collection and analysis of data that go beyond routine performance measurement. Program evaluations generally require designated budgets and are often conducted by outside experts to provide an independent assessment of program results.

A program evaluation should analyze performance measures to assess the achievement of performance goals, and, moreover, examine those achievements in the context of other aspects of program performance. Program evaluations may analyze relationships between program settings and services to learn how to improve program performance or to ascertain whether program activities have resulted in the intended outcomes. Some program evaluations attempt to isolate the influence of a program from other contributing factors, whereas performance measurement typically does not. Program evaluations can supplement performance reporting by measuring results that are too difficult or costly to assess routinely or by exploring why performance goals were not met.

Governance by performance

Performance management is pervasive in the federal and, indeed, at all levels of government. The public sector is expected to demonstrate its value and to seek new ways to foster performance. Public managers are asked to justify their actions in terms of efficiency and outcomes (**Moynihan, 2008, p. 3**).

Expectations for performance management in the federal government were codified in the Government Performance and Results Act (GPRA) enacted in 1993. Emphasis on sound evaluation and utilization of

findings has been reinforced by reauthorization of GPRA, as the GPRA Modernization Act of 2010. The act specifically directs federal officials to use performance information in decision making and holds them accountable for achieving results.

The Federal Performance Framework elaborated in Office of Management and Budget Circular A-11 (**OMB, 2013**) identifies three interrelated sets of activities: Planning; Evaluation, Analysis and Review; and Reporting. Performance management is further defined as:

> *"Use of goals, measurement, evaluation, analysis, and data-driven reviews to improve results of programs and the effectiveness and efficiency of agency operations. Performance management activities often consist of planning, goal setting, measuring, analyzing, reviewing, identifying performance improvement actions, reporting, implementing, and evaluating. The primary purpose of performance management is to improve performance and then to find lower cost ways to deliver effective programs."*

Performance management generates performance information through strategic planning, performance measurement, and program evaluation and connects this information to decision venues, where, ideally, the information influences a range of possible decisions.

The scope of NOAA's program evaluations

Program evaluation as a distinct field of professional practice is relatively new. Program evaluation in the United States began with the ambitious, federally funded social programs of the Great Society initiative during the mid- to late-1960s. Resources were spent on these programs, but the complex problems they were intended to address did not disappear. The public grew more cautious, and there was increasing pressure to provide evidence of the effectiveness of specific initiatives in order to allocate limited resources to ensure policies and interventions produce the desired outcome (**Kellogg, 1998, p. 4**).

> *"Program evaluation … was born of two lessons…: First, the realization that there is not enough money to do all the things that need doing; and second, even if there were enough money, it takes more than money to solve complex human and social problems. As not everything can be done, there must be a basis for deciding which things are worth doing. Enter evaluation."* (**Patton, 1997, p. 11**).

Program evaluation has since gained acceptance as a best practice in the public sector and is widely used by governmental and nongovernmental organizations alike. Federal agencies specialized in delivery of social services such as the US Agency for International Development, Center for Disease Control, and Department of Education, have well documented processes for program evaluation. NOAA programs are both broader and more specific than the commonly accepted domain of program evaluation, i.e., social services.

NOAA's mission requires a holistic view of environmental systems and must consider human activities in the context of environmental health and services. In contrast, NOAA managers are often called upon to decide between program design options and significant capital investments in earth observations, information technology, and restoration and recovery. Program evaluation as a method of systematic inquiry can both inform managerial decisions and communicate NOAA's value to the Nation. The general

principles of careful planning, diligent oversight, and objective analysis and reporting presented in this guide can improve the quality and utility of any evaluation irrespective of scope, scale or field of inquiry.

TYPES OF PROGRAM EVALUATIONS

We first explore different types of program evaluations, when it is appropriate to use them, and the benefits of each. Distinguishing evaluation type is often helpful in ensuring a program evaluation meets the needs of the requester. Virtually all program evaluations conducted in the federal government fall into four broad types: process, outcome, cost-benefit, and impact; which are discussed in turn below. In practice, a single program evaluation may address multiple research questions and multiple evaluation types may be employed in a single study.

Two factors should be considered when undertaking a program evaluation: the issue at hand and the stage of program development. Clarifying the issue is the key to deciding what methods to use and what type of evaluation is best suited to each research question. The stage of program development can limit the scope of available data and thus influence the evaluation type. Some common research questions asked at different stages of program development are listed in Table 1.

All program evaluations share common traits of rigorous planning, careful execution, thoughtful analysis, and thorough reporting. The steps described below for designing and implementing strong program evaluations apply regardless of evaluation type.

Table 1: Common research questions asked at different program stages.

Program stage	Common research questions	Evaluation type
Early stage of program or new initiative within a program	• Is the program being delivered as intended to the targeted recipients? • Is the program implemented as intended? • Have any feasibility or management problems emerged? • What progress has been made in implementing changes or new provisions?	Process
Mature, stable program with well-defined program model	• Are desired program outcomes obtained? • What, if any, unintended side effects did the program produce? • Do outcomes differ across program approaches, components, providers, or	Outcome

	client subgroups?	
	• Are program resources being used efficiently? • Why is a program no longer obtaining the desired level of outcomes?	Process
	• Did the program cause the desired impact? • Is one approach more effective than another in obtaining the desired outcomes?	Impact

Process Evaluations

Process evaluations, also called implementation evaluations, are the most frequently used type of evaluation. They review how a program is implemented and focus on how a program actually operates. In the context of a **logic model**, process questions address inputs, activities, and outputs.

Process evaluations can be beneficial throughout the life of a program, however they are often used when a program is implemented to ensure compliance with statutory and regulatory requirements, program design requirements, professional standards, and customer expectations. Early program evaluations can identify processes that can be made more efficient and mitigate compliance issues at a later date. A process evaluation may also be appropriate during the latter stages of a program life cycle when there is a need to assess program efficiency or effectiveness in achieving output goals.

Outcome Evaluations

Outcome evaluations, as the name implies, assess program outcomes. Thus, the focus is on the output-outcome portion of the **logic model**. Outcomes can be immediate effects of a program or more distal. In general, the closer the measure of an outcome is to program outputs, the clearer the linkage between the two. That is, outcomes measured immediately after outputs are generated are less likely to be affected by outside factors that can cloud the relationship between outputs and outcomes. A simple **scenario** is provided to illustrate the added complexity of measuring outcomes as they become more distal from the program.

In addition to intended outcomes, outcome evaluations should also address unintended outcomes. Referring to the **scenario** provided, skills gained through training may make trainees more attractive to other employers and result in higher turnover. Attention to the entire environment of a program is important to ensure significant contextual factors are not overlooked and competing influences on outcomes are considered.

Cost-Benefit Evaluations

Cost-benefit evaluations can be considered a special case of outcome evaluations. For these evaluations, program outputs and/or benefits are compared to input costs to provide a ratio of cost to benefit.

Because both costs and benefits are often difficult to assess, there can be substantial challenges implementing this analysis. Returning to the **scenario** provided, a survey may determine the extent to which trainees use their new skills, but assigning a value to skill use is not straightforward.

Impact Evaluations

Impact evaluations are designed to measure the net effect of a program by comparing actual program results with counterfactual data. Excluding all potential causes of an outcome can be a difficult, expensive proposition and is sometimes impossible. Because of their cost and required expertise, and often the need to plan the evaluation during initial program design rather than after program implementation, impact evaluations are not common. Although impact evaluations should be planned during program startup, they should not be undertaken until program operations are mature so that the true effect of the fully implemented program can be assessed.

The most straightforward way to isolate program impact is to randomly assign subjects (individuals, counties, cities, etc.) to treatment and control groups, i.e., groups that receive and do not receive program services. Experimental design is critical and factors such as group size and composition must be carefully considered to ensure a valid statistical sample. Treatment and control groups must also be sufficiently isolated to prevent spillover effects. This can be challenging since many programs cannot provide services selectively. An alternative to random assignment is to construct the control and treatment groups to be similar in ways that are considered important. Again, there is the issue of selective delivery of services and the additional issue of whether all important attributes have been considered in forming comparison groups. Advanced statistical techniques can be effective in isolating program effects in some cases; however, extensive data about the program and potential causal influences is often required.

PLANNING AND CONDUCTING AN EVALUATION

Program evaluations are typically completed in four stages: scoping, planning, implementation, and reporting. Program evaluations are generally team efforts and the questions posed below are provided to facilitate deliberations at each stage of the process.

Scoping

Many of NOAA's program evaluations are implemented by outside evaluators. Since external evaluators are not usually involved in scoping activities, it is especially important that program staff work through scoping activities to ensure any Request for Proposals specifies an appropriate type of program evaluation and the expectations for the evaluation are realistic. Whether planned and implemented internally or contracted, guidelines to ensure quality products apply.

Scoping activities should examine both the program itself and the purpose of the evaluation. While discussion questions addressing the program and purpose are listed separately below, these questions should be considered in tandem. The responses to these questions will likely change as the project progresses. It is nonetheless valuable to consider them fully in scoping to determine what evaluation questions are feasible, and how to get the best return on your investment.

Program Questions
Is the program "mature" enough to evaluate?
New starts and programs undergoing substantial change are generally not good candidates for comprehensive evaluation. Brief assessments of how an organization is managing a startup or major

change are always helpful to developing a strong program, but substantial program evaluation efforts should generally be reserved for mature and stable programs.

Is there a logic model or theory of change for the program?
A well-crafted **logic model** lays the foundation for selection of program evaluation methods and measures. For example, the activities described in the logic model, and their associated outputs, are central to development of process evaluation strategies. Outcomes, immediate and long-term, are the elements that drive measurement for outcome evaluations. If refinements of the logic model are needed, they should be developed during scoping to ensure the program is evaluable.

To what extent will an evaluation of the program contribute to strategic planning?
Authoritative research that informs NOAA's strategy should be reported and considered in agency strategic plans, budget narratives, and other planning documents. Thus, how the evaluation will contribute to strategic planning is an important criterion for deciding whether to proceed with an evaluation, and how resource-intensive it should be. Evidence from sound program evaluations should be a major driver in strategic decisions and should be documented in agency strategic plans, annual performance plan, and annual performance report (**OMB, 2013, section 210**). Strategic plans should also provide information about future evaluations and the rationale for selection of programs to evaluate.

To what extent is the program uniformly implemented?
Variations in program implementation between sites should drive key design decisions. Variation between sites can afford an opportunity to compare efficiency and effectiveness for different modes of implementation; however, it can also limit the information derived from the evaluation. Differences in the way data are defined and reported, for example, can limit methods of analysis and the conclusions that can be draw from available data. The criteria used to select evaluation sites should be carefully considered, particularly in deciding whether a subset of sites can be considered a representative sample and whether the results of an evaluation can be generalized.

What do past studies, monitoring data, expert opinion, management, staff, partners, and clients tell about program implementation and effectiveness?
Although some of the information from these sources may be anecdotal, potentially bias, or not entirely relevant to a potential program evaluation, information from multiple sources will help to complete a full picture of the program. Further, past studies may provide valuable information about the relationship between outputs and outcomes, as well as gaps in research the evaluation can fill. Performance monitoring data can identify gaps between expected and actual program performance, and views of those with knowledge of the program from various perspectives can inform what questions to address in an evaluation.

What data are already available?
Performance monitoring data, budget and other administrative data, case studies, and prior evaluations can establish baselines from which to measure progress. Evidence of outputs; program costs and staffing; changes over time in outputs, costs, and staffing add insights into evaluation design. Using existing data, where appropriate, can substantially reduce costs and time. The use of available data can have a major impact on the types of program evaluations that are feasible within cost and time constraints.

Are there mature performance measures in place that assess the history of program progress?

Although closely related to the prior question concerning available data, the importance of performance measurement information to developing sound program evaluations cannot be overemphasized. These measures may be key in planning and implementing outcome or impact evaluations. An assessment of the reliability and validity of these measures is vital to determining the types of evaluations that may be feasible.

Are there other programs or activities (federal, state, local, private) that are working with or coordinated with the program?
It is important to consider the context of the program: how it complements or conflicts with other programs, dependencies on other efforts, and the level of cooperation needed across programs to implement a credible evaluation. For impact evaluation, it is valuable even at this early stage to consider the extent to which program evaluation results are likely to identify the effects of the evaluated program itself apart from other related efforts. If the desired effects cannot be clearly attributed to the evaluated program, the value of an impact evaluation in relation to resources needed should be weighed.

Should partners be involved in the evaluation design and implementation?
Partners can often contribute resources to an evaluation. For example, partners may be willing to solicit cooperation in data collection, or, through past experience, may provide insights into study design, data collection, or methods of analysis.

Are there known external factors working against the program's ability to achieve its goals?
External factors can affect the appropriateness and timing of an evaluation, its potential uses, and its overall value. There are always external factors that hinder goal achievement, but the focus here is on significant events that disrupt or divert the program. For example, an economic downturn may have so severely affected the ability of local governments to actively address long-term environmental issues that evaluation would not represent the normal state of operations. Similarly, major events, such as Hurricane Katrina or Super Storm Sandy, likely put many programs into a different mode of operation. Evaluating a program immediately following such an extreme event could be misleading.

Purpose Questions
Who wants the evaluation and why?
Different stakeholders have different interests. If efficiency of operations is at issue, process questions may satisfy stakeholder needs. However, if the focus is the effect of the program, impact questions are required. The purpose(s) of the evaluation drive the research questions, which, in turn, drive the evaluation design, data collection, analysis, and reporting. There are often multiple drivers for an evaluation; and identifying the scope of interests will inform the needed scope of the evaluation. Clarifying the evaluation focus can also start to identify the level of precision needed. For program evaluations to satisfy mandates, legislative requirements should be carefully reviewed to ensure coverage.

How much time and what resources are available for the program evaluation?
Because program evaluations require careful design and, almost always, collection of new data, they can be both time consuming and costly. The research questions and study design must live within time and resource constraints. The availability of appropriate expertise for the various phases of the evaluation must be considered. Solid time and cost estimates for each stage of the program evaluation — planning, data collection, analysis, and reporting — are needed to ensure feasibility. Time estimates are often overly optimistic; the time required for data collection and analysis is frequently underestimated and

scheduling assumptions should be updated as new information becomes available. Generally, process evaluations are less costly and time consuming than impact evaluations because "process" generally confines data collection to the program itself (inputs, activities, and outputs) and impact evaluations require data beyond the program (outcomes).

What types of evaluation questions are of interest?
Although they overlap, process and outcome questions involve different evaluation strategies. Most evaluations include both process and outcome questions, but being specific about the types of questions can help gauge resource needs. Some research questions may not be feasible, and the sooner they are identified the better.

Must the results be generalizable?
If the research questions focus on acquiring detailed information about how programs operate, less emphasis may be centered on generalization as a primary requirement and more on case-study strategies. A careful selection of sites that either demonstrate the range of programs, or alternatively the most frequent type of program are examples of strategies that may be appropriate in situations where detail is of greater importance than the ability to generalize. Conversely, if the ability to generalize is essential, the level of detailed program information may need to be limited. A carefully crafted sampling strategy can frequently result in generalizable information without collecting data from the entire population.

How precise do the results need to be?
The intended use of information should be considered in deciding the required level of precision. Some cases have a greater tolerance for error than others. From a policy standpoint, an order of magnitude estimate of the number of students trained may be sufficient. This could allow local trainers to estimate the number of students rather than provide exact figures and thus streamline data collection. This could reduce data collection costs without compromising the usefulness of the evaluation. Other questions require the exact number of students; determining cost-per-student requires accurate estimates of both costs and number of students served, for example.

What is the best way to report evaluation results?
Even at this early stage it is important to consider the audience, how the information will be used, the level of detail, and deadlines for providing evaluation results. Adequate time for drafting and review of the report must be incorporated into the timeline for completing the project. If the plan is for multiple products, then additional time may be needed. Also, if time for reporting is constrained, consider ways to streamline delivering the evaluation findings (such as a briefing) followed by a written report to help relieve time constraints. Reporting plans can change over the life of the project, but consideration of reporting requirements in scoping can provide time to carefully consider the message before producing a report.

Planning
Time spent planning can result in fewer problems at the end of an evaluation. Potential research questions should be carefully screened. Well-crafted questions that are agreed to by all parties and meet the requirements of answerability go a long way toward a solid evaluation delivered on time and on budget. A **design matrix** is a valuable tool for ensuring a sound evaluation design and is a guide to evaluation execution, as well as to drafting the final report. A design matrix should be prepared as part of the planning process to clarify the researchable questions, identify sources of data, methodology,

analysis plans, limitations, and expectations about what will be reported. Proposals prepared by external evaluators should include a design matrix.

Planning should not be a step-by-step operation, but rather a process of considering each research question in relation to all other aspects of planning so that the final questions meet all tests of "research-ability." Time and resources for completing the evaluation must be ever-present in planning an efficient and effective evaluation. While the final plan should be achievable and documented, it is entirely appropriate to start by considering what questions would be answerable in an ideal situation, and stepping back from there to researchable questions. This helps to ensure that the evaluation is as informative and useful as possible within time, resource, and logistic constraints.

Program evaluations have much in common with projects: they are short term activities implemented to address specific issues. Like projects diligently monitoring cost, schedule and performance is crucial. Project planning tools can be helpful in organizing activities and scheduling resources. A simple spreadsheet specifying activities, time, and resources required for each phase of the evaluation can help ensure the evaluation is delivered as planned. The project plan should be revisited periodically and updated as needed.

Planning Questions

What are the researchable questions for the evaluation?
Often, the initial draft of researchable questions is not the final set. Research questions are often refined based on the underlying needs for the evaluation, the availability of information to address the questions, and time and resource constraints. Questions may focus on specific components of a program: What are the program's resources? What processes are used to accomplish program goals? Who is involved? What outputs have resulted? To what extent has the program achieved its goals? Clearly, for any given program there can be dozens of interesting questions. The researchable questions are those that are most important to achieving the evaluation's objectives and are answerable within the scope and resources available. Questions must address a single variable or relationship, be clear and concise, contain only well-defined terms, and be answerable. Questions should not imply an expected answer; they should be neutral. It is important to create a design strategy, data collection plan, and analysis plan that will address each question. For each researchable question, the type of question should be specified. Ask yourself, does each question address process, output, outcome, or impact?

What is the scope of the evaluation?
The program, timeframe, and locations must be defined and spelled out in specific terms. If there are multiple implementations of a program, the one that is the focus of the evaluation should be identified. If multiple configurations of a program (such as locally-designed activities that are all responses to a specific grant) are included in the evaluation, that should be specified and planned for as part of the evaluation design. In relation to time, an evaluation question may, for instance, address outcomes for program participants who *initially* enrolled in the program in 2010, or participants who were in the program *during any time* during 2010 (which could include those who initially started in 2009 as well as 2010), or only those who both *began and ended* the program in 2010, etc. Location could include all locations for the program, only locations that have been active for one year, five selected case study locations, locations along the coast of California, or any other specified subset.

What methodologies will be used to conduct the program evaluation?
The methodologies to employ are specifically linked to the researchable questions and the availability of data. It is likely that some questions within the same evaluation will be addressed using different methodologies, and that some questions may require the use of multiple methods to arrive at an answer. For example, questions about administrators' views of the program could be collected through phone interviews. Questions about how local government officials view the program may be collected through a web-based survey. Questions about how program applications are processed could be addressed by reviewing operating manuals, direct observations of applicant interviews, or staff interviews. The methods selected should be the most cost effective use of resources to generate the information required to address the question.

What information is required to address each question and what are the sources of information?
Both the information needed and the sources of that information must be clearly specified during planning. For example, if demographic data on trainees is going to be collected from trainee applications, then the specific demographics to be collected (age, sex, ethnicity, etc.) and the precise source of the information should be specified. For extant data or information (datasets, administrative records, etc.) the reliability and validity of the information for the purposes of the evaluation must be confirmed during the planning phase. If permission is required to collect needed data, it should be obtained during planning. If a survey is planned, the type of survey, sampling strategy, instrument design and likely response issues must all be considered. Sources of data should be free of bias as well as free of the appearance of bias. Bias or the potential for bias can undermine the credibility of findings.

What analysis plans will be used to assess the evaluation data?
Planning the analysis is essential to ensure that all needed data are collected, and that time is not wasted collecting information that is not used. Addressing questions such as unit of analysis (will program *participant* be the unit of analysis, or will *program* be the unit) will help to guide data collection to the most useful level. If content analysis is going to be used to analyze interview data, the planning stage is the time to develop an interview write up format, or computerized content analysis strategy to ease the eventual analysis process. Also, contingency plans for analysis should data collection problems arise — such as due to a low survey response rate — can mitigate the damage to the evaluation.

What significant questions will not be addressed by the evaluation?
Some evaluative questions may not be researchable with available resources. Such questions should be carefully considered to ensure opportunities to address them have not been overlooked. The audience for the evaluation should also understand the reasons for their omission.

What will the program evaluation allow you to say?
A clear statement of what the evaluation will support in relation to the research questions is essential to assessing whether the evaluation is worth the resources and time. For example, descriptive information on how a program operates at three sites may be sufficient to address many policy decision-making needs. However, if the evaluation will assess impact, descriptive information is probably not sufficient.

What resource and time commitments are needed to complete the evaluation?
Consideration of time and resource needs is integral to the planning process. If specific skills are needed at specific times, ensuring that right staff are available will help keep the evaluation on schedule. For example, if an expert is needed to develop a survey instrument and another expert is needed to draw the sample, then both experts must be available prior to distributing the survey to respondents. If you are planning a survey, time for pretests and required OMB clearances needs to be included. If site visits

are needed for data collection, make sure the appropriate staff will be available on site, and that travel expenses are within budget. Software and finding those who have the expertise to use it may also be crucial to completing the analysis of data. Budgeting time for staff leave and work on other assignments can help to make planning and reality match. A number of software packages are available to document activity, time, and resource plans.

Implementation

Careful planning is essential, but does not ensure a successful program evaluation: Low survey response rates, unforeseen data analysis problems, bad weather, unexpected changes in staffing, or a major change in how the program itself is implemented are just a few of the issues that can derail a well-planned evaluation. Successful implementation requires constant monitoring and the ability to make timely changes when needed.

Implementation Questions

How are the design matrix and the project management tool used in implementation?
Both the design matrix and the project management tool should be regularly compared to how the evaluation is actually progressing. Any differences should be considered an early warning of problems and addressed. Contingencies to resolve problems, which should have been developed in the planning phase, may need to be implemented. If there are substantive changes in the evaluation design or timeline, these changes should be documented so that these updated tools are in place to monitor further progress. Also, these updates can inform others about what to expect in the evaluation.

How often should the evaluation team meet during implementation?
While there are no hard and fast rules to govern how often a project team should meet during implementation, it is important everyone is informed of progress and problems. Evaluation team management, data collectors, analysts, and others are interdependent over the much of the project period. If issues arise and adjustments are needed, they will likely affect everyone involved in implementation, and solutions to problems could arise from any source: If data collection is taking longer than anticipated, those involved in collecting data may be best positioned to provide a solution. Delays in data collection could force schedule changes for data analysis and report writing. If the rate of data collection cannot be improved and the reporting deadline cannot be changed, analysis may be able to proceed with whatever data are already collected to get early results, and those early results could facilitate early drafting of the report message. Flexibility is often the key to a successful evaluation.

When should data be prepared for analysis?
Data should be prepared for analysis as soon as possible after collection. Interview write-ups should be completed as soon as feasible after each interview. Extant data should be tested to ensure it meets reliability and validity standards as soon as it is received. Survey responses should be coded as collected. Early tests of the analysis plans with preliminary data may pinpoint data collection issues that could be corrected before it is too late to change. Further, data coding schemes and analysis plans can be test-run on early data to determine if adjustments are needed.

When should analysis plans be modified?
Preliminary results often identify a need for additional analysis, and analysis plans should be modified accordingly to reflect additional major analyses. Unexpected findings should always be explored to better understand results and to confirm the adequacy of research methods. For example, a finding that new employees respond differently to formal training than long-term employees may recommend these two groups of employees be analyzed separately in relation to job satisfaction and attitudes toward on-

the-job training. Similarly, if response rates are lower than expected and fall below standards for generalization, the analysis may have to be confined to descriptive statistics rather than inferential statistics.

Reporting and Use
Planning for how evaluation results are reported should start during scoping activities. The purpose of the evaluation and the anticipated audience should drive the format, level of technical sophistication, and length of the report. In some cases, multiple reporting vehicles (e.g., a full-length report and a power point presentation) may be developed for different audiences.

Reporting Questions
How will the design matrix help develop the report message?
The researchable questions in the design matrix provide a start for organizing the report. Gathering information from all sources (past studies, survey results, interviews with program officials, etc.) in relation to each question will determine how definitively each question can be answered. The specific questions in the design matrix and their order do not always lend themselves to reporting, but they are a good start in organizing the data to develop the message.

How can the evaluation team decide on the message?
Team members may have different perspectives on what the data show. Holding a team meeting to share perspectives and agree on key messages can minimize changes once a draft is completed. The tone of the report should be discussed in addition to contents. Working as a team to identify the major points and then using that information to develop a detailed outline will help each team member understand the overall reporting strategy and relationships among the findings. If writing responsibilities are divided, all writers should understand the context of their assigned sections. Having someone who was not involved in the evaluation review the evidence can provide a valuable new perspective when evaluation team members cannot agree on the message.

Are graphics or text best for presenting findings?
Both graphics and text are helpful. Tabulated data can be used to provide detailed information in the body of the report (or in the appendices) and should be accompanied by text outlining the major points from the table. Graphs, pie charts, and other visual representations of findings are also shortcuts for readers to understand the message. All graphics and tables should be appropriately labeled including data sources, size of populations or samples as appropriate, and caveats. Text in short paragraphs or bullets, or broken up with graphic presentations is more user-friendly than long detailed paragraphs of dense text. Text should cover findings, conclusions, and recommendations in clear and concise language.

What should be covered in the body of the report and what should be in the appendices?
A message agreement meeting should include a discussion of what should be covered in the body of the report. Major findings should comprise the body of the report and appendices should contain information that is important, although not central to the message. For example, an overview of the methodology utilized in the evaluation should be in the report body, and details of methods can be in an appendix. Tables that are not essential to the message can be included in an appendix. Findings, other than major findings, that have been developed as a result of the evaluation should be included in appendices, not in the body of the report.

How long should the report be?
The report should be as short as possible while still conveying the message and appropriate caveats. Background information in the report should be the minimum necessary for the reader to understand the context of the evaluation. Methodology discussion in the report body should be brief, but any design or data issues should be noted so that the reader understands any limitations to the findings. Sufficient information from the evaluation should be presented to allow the reader to be able to track findings to conclusions and recommendations, and that is all.

Should the report have an executive summary?
An executive summary is essential to convey the message to those who may not read the entire report. The summary should focus on major findings, conclusions, and recommendations. It should be at a high level and distill the major messages of the evaluation. Generally, all subjects covered in the report need not to be included in the summary; however, new information should not be introduced in the summary.

Should the report include negative findings?
Negative findings are often the impetus for program changes and should be fully reported. If potential explanations for negative findings have been developed, they should be discussed. The program management response to negative findings should also be captured if practical.

Use Questions
What steps should be taken to foster use of the evaluation findings?
Evaluations are both expensive and infrequent; two good reasons to ensure evaluation results are used. Program evaluations should help program managers assess program performance and focus on changes that will improve performance. To foster use, evaluation reports should be shared with program staff, as well as stakeholders from other programs that share common goals. If an evaluation identifies areas for improvement in efficiency or effectiveness, an action plan can be helpful to identify practical steps towards positive change. It is also valuable to assess the evaluation itself: What worked well? What could have been done better? This critique can be helpful in planning and implementing a stronger evaluation the next time around. Sharing both the evaluation and the lessons learned with stakeholders can increase the value of the evaluation.

REFERENCES

Government Accountability Office. (2011). *Performance measurement and evaluation: Definitions and relationships.* Publication no. GAO-11-646SP. Retrieved from http://www.gao.gov/new.items/d11646sp.pdf

Moynihan, Donald P. (2008). *The dynamics of performance management: constructing information and reform*. Washington, DC: Georgetown University Press.

Office of Management and Budget. (2013). *Strategic plans, annual performance plans, performance reviews, and annual program performance reports (OMB Circular A11-Part 6).* Retrieved from http://www.whitehouse.gov/sites/default/files/omb/assets/a11_current_year/part6_executive_summary.pdf

Patton, Michael Q. (1997). *Utilization-Focused Evaluation, 3rd ed.,* Beverly Hills: Sage Publications.

W.K. Kellogg Foundation. (2010). *The W.K. Kellogg Foundation Evaluation Handbook.* Battle Creek, MI: W.K. Kellogg Foundation. Retrieved from http://www.wkkf.org/resource-directory/resource/2010/w-k-kellogg-foundation-evaluation-handbook

APPENDICES

Appendix 1: Performance Management and Program Evaluation

Performance measurement and program evaluation are related but distinct aspects of performance assessment. Performance measurement, as shown below, is a routine requirement for programs to assess progress towards goals. Program evaluations, in contrast, are specially planned research projects intended to provide findings and conclusions about the program as a whole that augment performance measurement information. Both forms of assessment support resource allocation and other policy decisions to improve service delivery and program effectiveness.

Performance Management vs. Program Evaluation

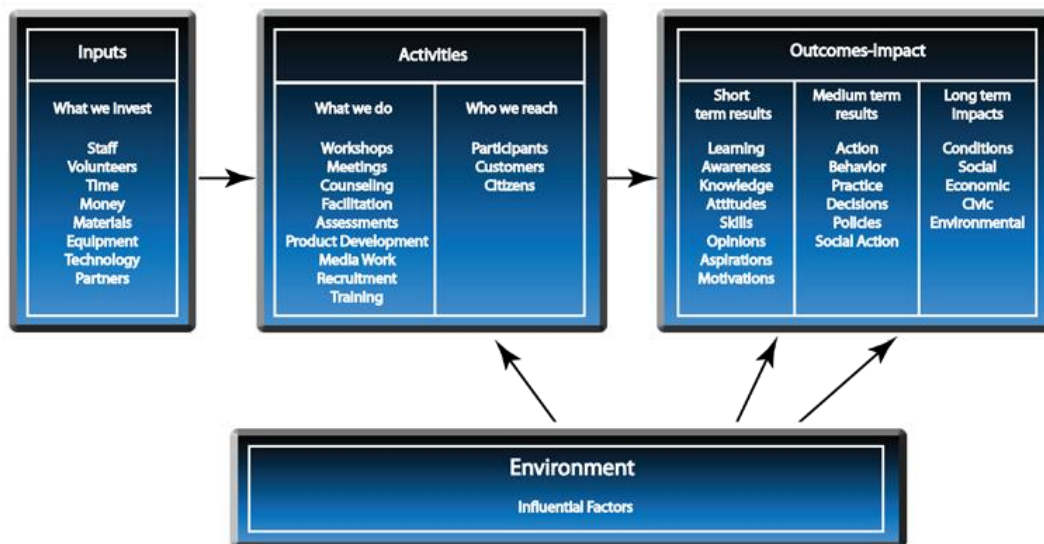| Dimension	Performance Measurement	Program Evaluation
Definition	**Ongoing monitoring and reporting** of program accomplishments, especially toward established goals.	**Individual systematic studies** conducted to assess program processes and/or impacts in detail.
Requirements	GPRA, OMB requirements, individual program mandates, agency policies.	GPRA-MA, OMB requirements, individual program mandates, agency discretion.
What gets measured	Progress towards goals, milestones, type and level of program activity, counts of products delivered, clients served, etc. Can include short-term outcomes.	Details of how the program operates (process evaluation), and/or outcomes/impact—intended and unintended--separated from the effects of other influences.
Use	Measures progress towards goals, and need for course corrections, provides accountability to the public.	Provide operational details that can lead to program efficiencies, determine extent to which program is affecting desired changes, and the relationship of costs to changes.
Who does it?	Generally, **program/agency staff.**	Generally, **external evaluation experts.**
Resource/time requirement	Planning and implementing collection can require substantial resources, but routine data collection is likely to require limited resources to compile and present data periodically.	Planning and implementation can be costly and time consuming. Program evaluations can require months to years of effort and can be high cost.
Frequency	**Periodic**—quarterly, yearly.	**Infrequent**, or, not at all.
Source of data	Program-generated documents, administrative records, limited new data collection.	Performance measurement sources **and** new data collection efforts—surveys, structured interviews, collection of extant data from program and elsewhere (such as census data).
Analysis	Counts assessed against targets, and/or compared to prior counts of same measure for evidence of progress.	Customized for the specific evaluation, can include sophisticated methods to isolate programs effects (impacts) from other influences.
Reporting	**Included in required periodic reporting** of program performance (GPRA, etc.).	**Separate evaluation report**, and summarized in agency planning documents.
Cost	Generally **absorbed into overall program budget.**	Generally requires **added funding.**
Paperwork Reduction Act (for surveys)	OMB approval needed for data collection from more than 9 respondents.	OMB approval needed for data collection from more than 9 respondents.
Training needed	Data collection and recording methods, understanding of measurement issues, basis data analysis	Evaluation design and implementation, basic and advanced data analysis, managing evaluation contractors.

Appendix 2: Logic Models

A logic model is a common tool used to understand how a program should operate and how it actually operates. Evaluators use logic models to explain the strategy, or logic, by which a program is expected to achieve its goals. A logic model is typically represented as a flow chart that tracks how inputs drive activities to produce outputs, outcomes, and the ultimate goal of the program. By specifying expectations at each step of a program, a logic model can help evaluators articulate the assumptions and expectations of program managers and stakeholders. By specifying expectations, a logic model can help evaluators define measures of the program's performance and track progress toward its ultimate goal. A variety of formats can be used; the key is to develop a clear understanding of the program, the context in which it operates, and the policy issues involved.

A logic model can be helpful as a:

- Program planning tool: depicting the implications for program design of previous research on the key factors influencing achievement of the desired benefits;

- Communication tool: encouraging shared understanding and expectations among policy makers and program managers and obtaining the support and cooperation of program partners;

- Program implementation tool: mapping what activities should occur at various times and which groups should be involved; and defining performance measures and evaluation questions.

A logic model for public engagement

In describing a program's goals and strategies, it is important to consult a variety of sources—legislative history, program staff and materials, prior research on the program, public media, and congressional staff—to uncover differences in expectations and concerns stakeholders may have. It is also important to understand the context, i.e., why the program was initiated; whether circumstances have changed since its inception; and current policy concerns. In the absence of clearly established definitions of the intervention or its desired outcomes, the evaluator will need to discuss these issues with the evaluation requestor and may need to explore, as part of the evaluation, how the program and its goals have been operationally defined.

Appendix 3: A Program Evaluation Scenario

An intensive course is provided to teach specific skills to first responders. The goals of the program are that trainees learn new skills, utilize their new skills on the job, and transfer new skills to coworkers through informal on the job training. Trainees are tested at the end of the course and are expected to achieve a score of 80 percent or better to pass the course. No prior knowledge is assumed.

Table 3: Program Evaluation Scenario

Type of Outcome	Outcome	Method of Assessment	Potential Confounding Factors
Immediate	New skills learned	Final exam	Trainees may receive additional training from other sources.
Intermediate	New skills used	Survey trainees about use of new skills	Trainees may have been assigned to tasks where their new skills were not used. Survey responses may be inaccurate.
Long-term	New skills transferred to coworkers	Survey supervisors to assess transfer of skills from trainees to others	Trainees are no longer in the work group. Coworkers receive training from other sources. New, superior skills supplant trainee skills and new training is required. Survey responses may be inaccurate.

Appendix 4: Design Matrix

The design matrix is a standard tool used to outline the components of an evaluation design, as well as the limitations of design choices. A design matrix should be completed for each significant project to document design decisions and summarize the key issues in the evaluation design. All staff having significant involvement in or oversight of the work meet to discuss this plan and reach agreement on whether it can credibly answer the evaluation questions. The design matrix can also be used to document and ensure compliance with program evaluation requirements.

Guidance for the design matrix is shown below to demonstrate the issues, design choices, and trade-offs that an evaluator should consider. The guidance is fairly general but challenges the evaluator to justify the design components for each researchable question. Finally, the tool can help stakeholders understand the logic of the evaluation.

Design matrix template.

Issue Statement: Provide a few sentences about the program and the purpose of the evaluation.

Researchable Questions	Information Required /Information Sources	Scope and Methodology	Limitations	What this analysis will likely allow you to say
What questions is the team trying to answer? Identify specific questions that must be addressed to satisfy the objectives of the evaluation. Ensure each major evaluation question is specific, objective, neutral, measurable, and doable. Ensure key terms are defined. Each major evaluation question should be addressed in a separate row.	**What information is needed to address each evaluation question? Where will they get it?** Identify documents or types of required information. Identify plans to address internal controls and compliance. Identify plans to collect documents that establish the "criteria" to be used. Identify plans to follow up on known significant findings and open recommendations found in obtaining background information. Identify sources of the required information, such as databases, studies, subject area	**How will the team answer each evaluation question?** Describe strategies for collecting the required information or data, such as random sampling, case studies, focus groups, questionnaires, benchmarking to best practices, use of existing data bases, etc. Describe the planned scope of each strategy, including the timeframe, locations to visit, and sample sizes. Describe the analytical techniques to be used, such as regression analysis, cost benefit analysis, sensitivity analysis, modeling, descriptive analysis, content analysis, case study summaries, etc.	**What are the design limitations and how will they affect the end product?** Cite any limitations as a result of the information required or the scope and methodology, such as: —Questionable data quality and/or reliability. —Inability to access certain types of data or obtain data covering a certain time frame. —Security classification or confidentiality restrictions. —Inability to generalize or extrapolate findings. Be sure to address	**What are the expected results of the work?** Describe what you can likely say. Draw on preliminary results for illustrative purposes, if helpful. Ensure the proposed answer addresses the evaluation question identified in column one.

	experts, program officials, models, etc.		how these limitations will affect the product.	