

## Tip Sheet 4: Information on and Resources for Designing Evaluations

Step 3 of this guide states that analysts should conduct a new evaluation if existing evaluations are not available, relevant, or sound. Evaluations can help analysts validate the effects of fragmentation, overlap, or duplication or assess and compare the performance of programs. This tip sheet provides resources on how to scope, design, and conduct an evaluation.

Evaluations are studies tailored to answer specific questions about how well a program is working. Program evaluation is closely related to performance measurement and reporting. Program evaluations typically examine a broad range of information on program performance, whereas performance measurement is the systematic ongoing monitoring and reporting of program accomplishments, particularly progress toward pre-established goals or standards.<sup>31</sup> Performance measures or indicators may address program staffing and resources (or inputs), the type or level of program activities conducted (or process), the direct goods or services delivered by a program (or outputs), or the results of those goods and services (or outcomes). GAO has issued several products on designing effective program evaluations (see the list of key GAO products at the end of this tip sheet).

GAO has outlined five steps analysts should take when designing evaluations:

1. Clarify understanding of the program's goals and strategy.
2. Develop relevant and useful evaluation questions.
3. Select an appropriate approach or design for each evaluation question.
4. Identify data sources and collection procedures to obtain relevant, credible information.
5. Develop plans to analyze the data in ways that allow valid conclusions to be drawn from the evaluation questions.

### Defining the Evaluation's Scope and Selecting an Evaluation Design

Because an evaluation can take a number of directions, the first steps in its design aim to define its purpose and scope—to establish what questions it will and will not address. The evaluation's scope is tied to its research questions and defines the subject matter it will assess, such as a program or aspect of a program, and the time periods and locations that will be included. To ensure the evaluation's credibility and relevance to its intended users, **the analyst must develop a clear understanding of the program's purpose and goals and develop researchable evaluation questions** that are feasible and appropriate to the program and that address the intended users' needs.

Once evaluation questions have been formulated, the next step is to **develop an evaluation design**—to select appropriate measures and comparisons that will permit drawing valid conclusions on those questions. In the design process, the analyst explores the variety of **options available for collecting and analyzing information and chooses alternatives that will best address the evaluation objectives within available resources**. Selecting an appropriate and feasible design, however, is an iterative process and may result in the need to revise the evaluation questions.

The basic components of an evaluation design include the following:

- the evaluation questions, objectives, and scope;

<sup>31</sup>More specifically, performance measurement focuses on whether a program has achieved its objectives, expressed as measurable performance standards. Program evaluations typically examine a broader range of information on program performance and its context than is feasible to monitor on an ongoing basis. In addition, both forms of assessment aim to support resource allocation and other policy decisions to improve service delivery and program effectiveness. But performance measurement, because of its ongoing nature, can serve as an early warning system to management and as a vehicle for improving accountability to the public.

- information sources and measures, or what information is needed;
- data collection methods, including any sampling procedures, or how information or evidence will be obtained;
- an analysis plan, including evaluative criteria or comparisons, or how or on what basis program performance will be judged or evaluated; and
- an assessment of study limitations.

Strong evaluations employ methods of analysis that are appropriate to the question; support the answer with sufficient and appropriate evidence; document the assumptions, procedures, and modes of analysis; and rule out competing explanations. Strong studies present questions clearly, address them appropriately, and draw inferences commensurate with the power of the design and the availability, validity, and reliability of the data. Thus, a good evaluation design should do the following:

- **Be appropriate for the evaluation questions and context.** The design should address all key questions, clearly state any limitations in scope, and be appropriate for the nature and significance of the program or issue. For example, evaluations should not attempt to measure outcomes before a program has been in place long enough to be able to produce them.
- **Adequately address the evaluation question.** The strength of the design should match the precision, completeness, and conclusiveness of the information needed to answer the questions and meet the analyst's and decision makers' needs. Criteria and measures should be narrowly tailored, and comparisons should be selected to support valid conclusions and rule out alternative explanations.
- **Fit available time and resources.** Time and costs are constraints that shape the scope of the evaluation questions and the range of activities that can help answer them. Producing information with an understanding of the user's timetable enhances its usefulness, but limitations and constraints of the evaluation must be disclosed.
- **Rely on sufficient, credible data.** No data collection and maintenance process is free of error, but the data should be sufficiently free of bias or other significant errors that could lead to inaccurate conclusions. Measures should reflect the persons, activities, or conditions that the program is expected to affect and should not be unduly influenced by factors outside of the program's control.

### Designs for Assessing Program Implementation and Effectiveness

Program evaluation designs are tailored to the nature of the program and the questions being asked. Thus, they can have an infinite variety of forms as analysts choose performance goals and measures and select procedures for data collection and analysis. Nevertheless, individual designs tend to be adaptations of a set of familiar evaluation approaches—that is, evaluation questions and research methods for answering them. The following tables provide examples of some typical evaluation approaches for implementation and effectiveness questions and examples of designs specifically matched to program structure.

Implementation (or process) evaluations address questions about how and to what extent activities have been implemented as intended and whether they are targeted to appropriate populations or problems. Table 10 provides examples of implementation questions and designs used to address them.

**Table 10: Common Designs for Implementation (or Process) Evaluations**

Evaluation question	Design
Is the program being implemented as intended?	Compare program activities to statute and regulations, program logic model, professional standards, or stakeholder expectations.
Have any feasibility or management problems emerged?	<p>Compare program performance to quality, cost, or efficiency expectations.</p> <p>Assess variation in quality or performance across settings, providers, or subgroups of recipients.</p>
Why is the program no longer achieving expected outcomes?	<p>Analyze program and external factors correlated with variation in program outcomes.</p> <p>Interview key informants about possible explanations.</p> <p>Conduct in-depth analysis of critical cases.</p>

Source: GAO. | GAO-15-49SP

Outcome evaluations address questions about the extent to which the program achieved its results-oriented objectives. This form of evaluation focuses on examining outputs (goods and services delivered by a program) and outcomes (the results of those goods and services) but may also assess program processes to understand how those outcomes are produced. Table 11 provides examples of outcome-oriented evaluation questions and designs to address them.

**Table 11: Common Designs for Outcome Evaluations**

Evaluation question	Design
Is the program achieving its desired outcomes or having other important side effects?	<p>Compare program performance to law and regulations, program logic model, professional standards, or stakeholder expectations.</p> <p>Assess change in outcomes for participants before and after exposure to the program.</p> <p>Assess differences in outcomes between program participants and nonparticipants.</p>
Do program outcomes differ across program components, providers, or recipients?	Assess variation in outcomes (or change in outcomes) across approaches, settings, providers, or subgroups of recipients.

Source: GAO. | GAO-15-49SP

Many desired outcomes of federal programs are influenced by external factors, including other federal, state, and local programs and policies, as well as economic and environmental conditions. Thus, the outcomes observed typically reflect a combination of influences. To isolate the program’s unique impacts, or contribution to those outcomes, an impact study must be carefully designed to rule out plausible alternative explanations for the results. Table 12 provides examples of designs commonly used to address questions related to causal inferences about program impacts.

**Table 12: Common Designs for Drawing Causal Inferences about Program Impacts**

Evaluation question	Design
Is the program responsible for (effective in) achieving improvements in desired outcomes?	<p>Compare (change in) outcomes for a randomly assigned treatment group and a nonparticipating control group (randomized controlled experiment).</p> <p>Compare (change in) outcomes for program participants and a comparison group closely matched to them on key characteristics (comparison group quasi-experiment).</p> <p>Compare (change in) outcomes for participants before and after the intervention, over multiple points in time with statistical controls (single group quasi-experiment).</p>
How does the effectiveness of the program approach compare with other strategies for achieving the same outcomes?	<p>Compare (change in) outcomes for groups randomly assigned to different treatments (randomized controlled experiment).</p> <p>Compare (change in) outcomes for comparison groups closely matched on key characteristics (comparison group quasi-experiment).</p>

Source: Adapted from Bernholz et al., 2006. | GAO-15-49SP

### Selecting a Design

As evaluation designs are tailored to the nature of the program and the questions asked, it becomes apparent that certain designs are necessarily excluded for certain types of programs. Some types of federal programs, such as those funding basic research projects or the development of statistical information, are not expected to have readily measurable effects on their environment. Therefore, research programs have been evaluated on the quality of their processes and products and relevance to their customers' needs, typically through expert peer review of portfolios of completed research projects. Regulatory and law enforcement programs can be evaluated according to the level of compliance with the pertinent rule or achievement of desired health or safety conditions, obtained through ongoing outcome monitoring. Experimental and quasi-experimental impact studies are better suited for programs conducted on a small scale at selected locations, where program conditions can be carefully controlled, rather than at the national level. Such designs are particularly appropriate for demonstration programs testing new approaches or initiatives, and are not well suited for mature, universally available programs.<sup>32</sup> Table 13 summarizes the features of designs discussed above as well as the types of programs employing them.

<sup>32</sup>For more information on these design approaches, see GAO, *Designing Evaluations: 2012 Revision*, [GAO-12-208G](#) (Washington, D.C.: January 2012).

**Table 13: Designs for Assessing Effectiveness of Different Types of Programs**

Typical design	Comparison controlling for alternative explanations	Best suited for
Process and outcome monitoring or evaluation	<p>Performance and pre-existing goals or standards such as:</p> <ul style="list-style-type: none"> <li>• Research and design criteria of relevance, quality, and performance</li> <li>• productivity, cost-effectiveness, and efficiency standards</li> <li>• customer expectations or industry benchmarks</li> </ul>	<p>Research, enforcement, information and statistical programs, business-like enterprises, and mature ongoing programs where</p> <ul style="list-style-type: none"> <li>• coverage is national and complete</li> <li>• few, if any, alternatives explain observed outcomes</li> </ul>
Quasi-experiments: single group	<p>Outcomes for program participants before and after the intervention:</p> <ul style="list-style-type: none"> <li>• collects outcomes data at multiple points in time</li> <li>• statistical adjustments or modeling control for alternative causal explanations</li> </ul>	<p>Regulatory and other programs where</p> <ul style="list-style-type: none"> <li>• clearly defined interventions have distinct starting times</li> <li>• coverage is national and complete</li> <li>• randomly assigning participants is NOT feasible, practical, or ethical</li> </ul>
Quasi-experiments: comparison group	<p>Outcomes for program participants and a comparison group closely matched to them on key characteristics:</p> <ul style="list-style-type: none"> <li>• key characteristics are plausible alternative explanations for a difference in outcomes</li> <li>• measures outcomes before and after the intervention (pretest, post-test)</li> </ul>	<p>Service and other programs where</p> <ul style="list-style-type: none"> <li>• clearly defined interventions can be standardized and controlled</li> <li>• coverage is limited</li> <li>• randomly assigning participants is NOT feasible, practical, or ethical</li> </ul>
Randomized experiments: control groups	<p>Outcomes for a randomly assigned treatment group and a nonparticipating control group:</p> <ul style="list-style-type: none"> <li>• measures outcomes preferably before and after the intervention (pretest, post-test)</li> </ul>	<p>Service and other program where</p> <ul style="list-style-type: none"> <li>• clearly defined interventions can be standardized and controlled</li> <li>• coverage is limited</li> <li>• randomly assigning participants is feasible and ethical</li> </ul>

Source: Adapted from Bernholz et al., 2006. | GAO-15-49SP

### Design Approaches for Selected Methodological Challenges

The designs outlined previously may have limited relevance and credibility on their own for assessing the effects of federal programs where neither the intervention nor the desired outcome is clearly defined or measured. In addition, many, if not most, federal programs aim to improve some aspect of complex systems, such as the economy or the environment, over which they have limited control, or share responsibilities with other agencies for achieving their objectives. Thus, it can be difficult to confidently attribute a causal connection between the program and the observed outcomes. Federal agencies have implemented a number of strategies to address evaluation challenges and develop performance information for these types of programs that can inform management, oversight, and policy.

- **Challenge: Lack of common outcome measures.** A federal program might lack common national data on a desired outcome because the program is relatively new, new to measuring outcomes, or

had limited control over how service providers collect and store information. Where state programs operate without much federal direction, outcome data are often not comparable across the states. Federal agencies have taken different approaches to obtaining common national outcome data, depending in part on whether such information is needed on a recurring basis: (1) collaborating with others on a common reporting format, (2) recoding state data into a common format, and (3) conducting a special survey to obtain nation-wide data.

- **Challenge: Desired outcomes are infrequently observed.** Some federal programs are created to respond to national concerns, such as increased cancer rates or environmental degradation, which operate in a lengthy time frame and are not expected to be resolved quickly. Thus, changes in intended long-term outcomes are unlikely to be observed within an annual performance reporting cycle or even, perhaps, within a 5-year evaluation study. Other programs aim to prevent or provide protection from events that are very infrequent and, most importantly, not predictable, such as storms or terrorist attacks, for which it is impractical to set annual or other relatively short-term goals. Evaluation approaches to these types of programs may rely heavily on well-articulated program logic models to depict the program's activities as multistep strategies for achieving its goals. Depending on how infrequent or unexpected opportunities may be to observe the desired outcome, an analyst might choose to (1) measure program effects on short-term or intermediate goals, (2) assess the quality of an agency's prevention or risk management plans, or (3) conduct a thorough after-action or critical-incident review of any incidents that do occur.
- **Challenge: Benefits of research programs are difficult to predict.** The increased interest in assuring accountability for the value of government expenditures has been accompanied by increased efforts to demonstrate and quantify the value of public investments in scientific research. An analyst might readily measure the effectiveness of an applied research program by whether it met its goal to improve the quality, precision, or efficiency of tools or processes. However, basic research programs do not usually have such immediate, concrete goals. Instead, goals for federal research programs can include advancing knowledge in a field and building capacity for future advances through developing useful tools or supporting the scientific community. In addition, multiyear investments in basic research might be expected to lead to innovations in technology that will (eventually) yield social or financial value, such as energy savings or security. Common agency approaches to evaluating research programs include (1) external expert review of a research portfolio and (2) bibliometric analyses of research citations and patents.
- **Challenge: Benefits of flexible grant programs are difficult to summarize.** Federal grant programs vary greatly as to whether they have performance objectives or a common set of activities across grantees such as state and local agencies or nonprofit service providers. Where a grant program represents a discrete program with a narrow set of activities and performance-related objectives, such as a food delivery program for seniors, it can be evaluated with the methods under Selecting a Design. However, a formula or "block" grant, with loosely defined objectives that simply adds to a stream of funds supporting ongoing state or local programs, presents a significant challenge to efforts to portray the results of the federal or national program. Agencies have deployed a few distinct approaches, often in combination: (1) describe national variation in local approaches, (2) measure national improvement in common outputs or outcomes, and (3) conduct effectiveness evaluations in a sample of states.
- **Challenge: Assess the progress and results from comprehensive reforms.** In contrast to programs that support a particular set of activities aimed at achieving a specified objective, some comprehensive reform initiatives may call for collective, coordinated actions in communities in multiple areas, such as altering public policy, improving service practice, or engaging the public to create system reform. This poses challenges to the analyst in identifying the nature of the intervention (or program) and the desired outcomes, as well as an estimate of what would have occurred in the absence of these reforms. Depending on the extent to which the dimensions of reform are well understood, the progress of reforms might be measured quantitatively in a survey or through a more exploratory form of case study.

- **Challenge: Isolating impact when several programs are aimed at the same outcome.** Attributing observed changes in desired outcomes to the effect of a program requires ruling out other plausible explanations for those changes. For example, environmental factors such as historical trends in community attitudes towards smoking, could explain changes in youths' smoking rates over time. Other programs funded with private, state, or other federal funds may also strive for goals similar to those of the program being evaluated. Although random assignment of individuals to treatment and comparison groups is intended to cancel out the influence of those factors, in practice, the presence of these other factors may still blur the effect of the program of interest or randomization may simply not be feasible. Collecting additional data can help strengthen conclusions about an intervention's impact from both randomized and nonrandomized designs. In general, to help isolate the impact of programs aimed at the same goal, it can be useful to construct a logic model for each program—carefully specifying the programs' distinct target audiences and expected short-term outcomes—and assess the extent to which the programs actually operate in the same localities and reach the same populations. Then the analyst can devise a data collection approach or set of comparisons that could isolate the effects of the distinct programs, such as (1) narrowing the scope of the outcome measure, (2) measuring additional outcomes not expected to change, or (3) testing hypothesized relationships between the programs.

### Key GAO Reports

*Designing Evaluations: 2012 Revision.* [GAO-12-208G](#). Washington, D.C.: January 2012.

*Performance Measurement and Evaluation: Definitions and Relationships.* [GAO-11-646SP](#). Washington, D.C.: May 2011.

*Program Evaluation: A Variety of Rigorous Methods Can Help Identify Effective Interventions.* [GAO-10-30](#). Washington, D.C.: November 23, 2009.

### Other Key Resources

Office of Management and Budget, Circular No. A-11, *Preparation, Submission, and Execution of the Budget*, pt 6 (July 2014).