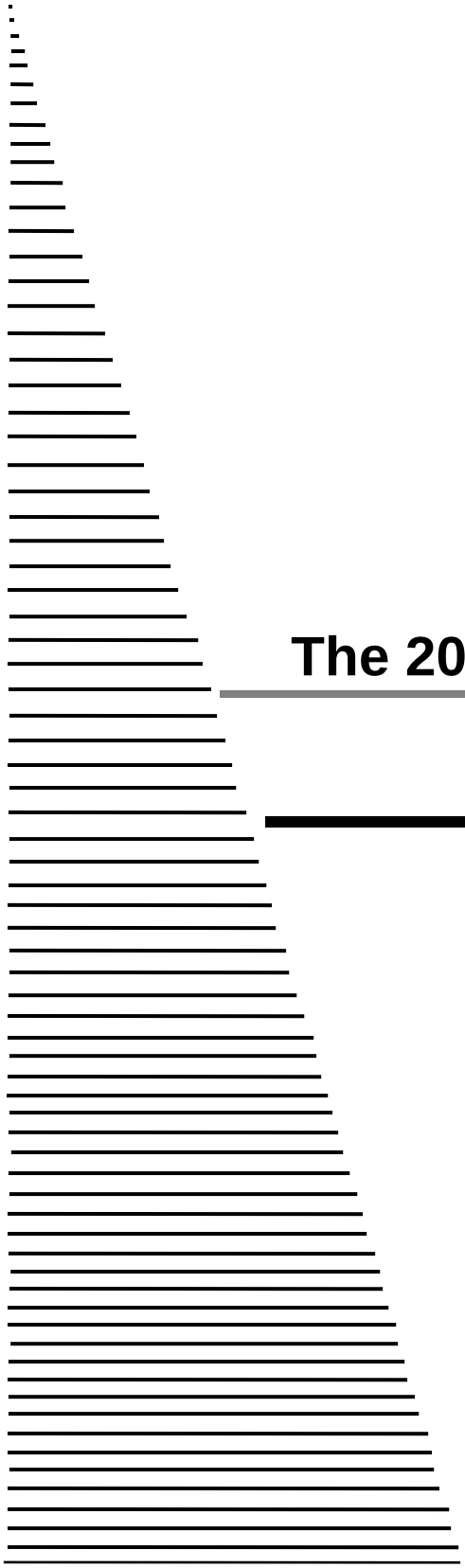
**Directorate for Education and Human Resources**


**Division of Research and Learning in Formal and Informal Settings**
**National Science Foundation**



# The 2010 User-Friendly Handbook for Project Evaluation

# The 2010 User-Friendly Handbook for Project Evaluation

Prepared under Contract
REC 99-12175

by

Joy Frechtling
Westat

with contributing authors of three special sections:

Melvin M. Mark

Debra J. Rog

Veronica Thomas, Henry Frierson, Stafford Hood, & Gerunda Hughes

Elmima Johnson
Program Officer and COTR
Division of Research and Learning in Formal and Informal Settings
National Science Foundation

December 2010

# TABLE OF CONTENTS

# TABLE OF CONTENTS (CONTINUED)

# TABLE OF CONTENTS (CONTINUED)

# TABLE OF CONTENTS (CONTINUED)

## List of Exhibits

# INTRODUCTION

This Handbook was developed to provide project directors and principal investigators working with the National Science Foundation (NSF) with a basic guide for evaluating NSF's educational projects. It is aimed at people who need to learn more about both the value of evaluation and how to design and carry out an evaluation, rather than those who already have a solid base of experience in the field. It builds on firmly established principles, blending technical knowledge and common sense to meet the special needs of NSF and its stakeholders.

The Handbook discusses quantitative and qualitative evaluation methods, suggesting ways in which they can be used as complements in an evaluation strategy. As a result of reading this Handbook, it is expected that principal investigators will increase their understanding of the evaluation process and NSF's requirements for evaluation, as well as gain knowledge that will help them to communicate with evaluators and obtain data that help them improve their work.

To develop this Handbook, we have drawn on the similar handbooks and tools developed for the National Science Foundation (especially the 1993 *User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education*, the 1997 *User-Friendly Handbook for Mixed-Method Evaluations*, and the *2002 User-Friendly Handbook for Project Evaluation*) and the National Aeronautics and Space Administration. However, special attention has been given to aligning the Handbook to NSF's unique needs and experiences. In addition, a range of NSF program areas have been selected to provide concrete examples of the evaluation issues discussed. The Handbook is divided into nine chapters:

Chapters 1 through 7 are updates of material included in earlier Handbooks. Chapters 8 and 9 are new additions to this Handbook focusing on rigorous project evaluation and the factors that contribute to it.

We have also provided a glossary of commonly used terms as well as references for those who might wish to pursue some additional readings. Appendix A presents some tips for finding an evaluator.

**References**

Frechtling, J., Stevens, F., Lawrenz, F., and Sharp, L. (1993). *The User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education*. NSF 93-152. Arlington, VA: NSF.

Frechtling, J., and Sharp, L. (1997). *The User-Friendly Handbook for Mixed-Method Evaluations*. NSF 97-153. Arlington, VA: NSF.

Frechtling, J. (2002) *The 2002 User-Friendly Handbook for Project Evaluation.* NSF 02-057. Arlington, VA: NSF.

Chapter 1

# REASONS FOR CONDUCTING EVALUATIONS

The notion of evaluation has been around for a long time. In fact, the Chinese had a large functional evaluation system in place for their civil servants as long ago as 2000 B.C. There are also various definitions of evaluation; some view it as tests, others as descriptions, documents, or even management. A comprehensive definition, as presented by the Joint Committee on Standards for Educational Evaluation (1994), holds that evaluation is "systematic investigation of the worth or merit of an object." This definition centers on the goal of using evaluation for a purpose. Accordingly, evaluations should be conducted for action-related reasons, and the information provided should facilitate some specific course of action.

*Evaluation provides information to help improve a project.*

Why should NSF grantees conduct evaluations? There are two very important answers to this question. First, evaluation produces information that can be used to improve the project. Information on how different aspects of a project are working and the extent to which the objectives are being met are essential to a continuous improvement process. Second, an evaluation can document what has been achieved. This aspect of the evaluation typically assesses the extent to which goals are reached and desired impacts are attained. In addition, and equally important, evaluation frequently provides new insights or new information that was not anticipated. What are frequently called "unanticipated consequences" of a program can be among the most useful outcomes of the assessment enterprise.

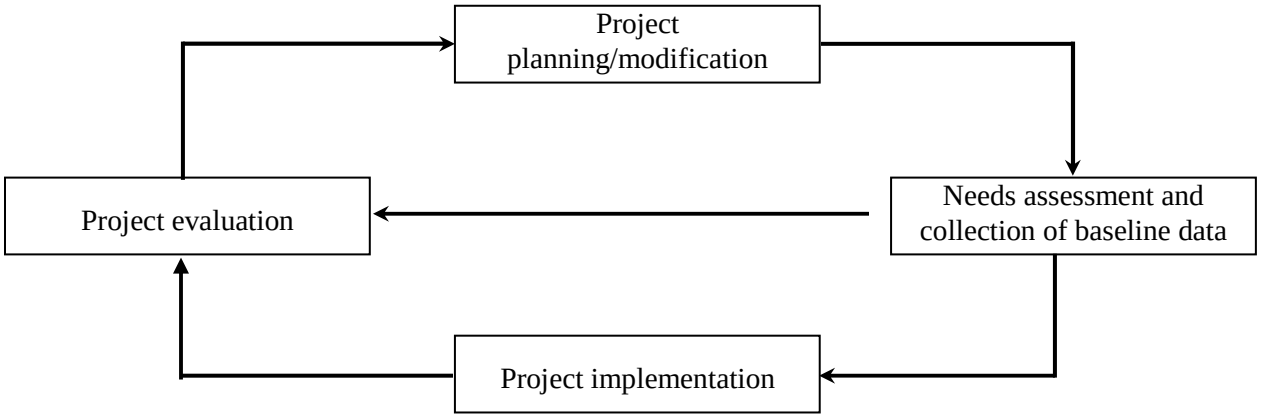
*Evaluations need not be conducted in an adversarial mode.*

Too frequently evaluation has been viewed as an adversarial process. In such cases, its main use has been to provide a "thumbs up" or "thumbs down" about a program or project. Hence, it has all too often been considered by program or project directors and coordinators as an external imposition that is threatening, disruptive, and not very helpful to project staff. While that may be true in some situations, evaluations need not be, and most often are not, conducted in an adversarial mode. Rather, they can contribute to the knowledge base to help understand what works and why.

The current view of evaluation stresses the inherent interrelationships between evaluation and program implementation. Evaluation is a valuable source of information on how the project is being implemented, specifically, what works and what should be modified. Furthermore, in contrast to the outdated belief held by some that evaluation should take place at the end of a project, the accepted wisdom is to incorporate it at the beginning of a project. Planning, evaluation, and implementation are all parts of a whole, and they work best when they work together. Kaser et al. (1999) go so far as to state that "a quality program takes evaluation seriously and builds it into the program design" (p. 23). Exhibit 1 shows

the interaction between evaluation and other aspects of your NSF project as the project is developed and initiated.

**Exhibit 1.—The project development/evaluation cycle**




*Evaluation provides information for communicating to a variety of stakeholders.*

Additionally, evaluation provides information for communicating to a variety of stakeholders. It allows project managers to better tell their story and prove the worth of their projects. It also gives managers the data they need to report "up the line," to inform senior decisionmakers about the outcomes of their investments. EHR's past evaluation efforts have been responsive to federal reporting requirements including the Government Performance and Results Act (GPRA) and the Office of Management and Budget's Program Assessment Rating Tool (PART). GPRA required federal agencies to report annually on the accomplishments of their funded efforts, reporting the results or impacts of the federal government through the establishment of broad goals or strategic outcomes, performance outcomes, and performance indicators. PART provided a systematic method for assessing the performance of federal program activities through a review of program purposes and design, strategic planning, management, and results and accountability. NSF efforts also incorporated the Academic Competitiveness Council's (ACC) recommendations in its assessment and accountability framework for science, technology, engineering, and mathematics (STEM) education.

The current administration's focus is on evaluation for problem solving and designing programs. NSF evaluation activities are providing data to inform the OMB call for high priority performance goals (HPPG). This process will involve each agency setting three to eight priority goals; identifying a goal leader for each goal; and creating an action plan to identify problems and solutions. A quarterly update on progress is planned, followed by an annual OMB performance review with results posted on a new website. The goal is to use performance information to improve outcomes, communicate results thereby improving transparency, and building a STEM education knowledge base.

**References**

Joint Committee on the Standards for Educational Evaluation. (2010). *The Program Evaluation Standards*. 2nd Ed. Thousand Oaks, CA: Sage Publications.

Kaser, J., Bourexis, P., Loucks-Horsley, S., and Raizen, S. (1999). *Enhancing Program Quality in Science and Mathematics*. Thousand Oaks, CA: Corwin Press.

Chapter 2

# EVALUATION PROTOTYPES

The purpose of this chapter is to provide a grounding in evaluation and to discuss the kinds of information evaluation can provide. We start with the assumption that the term "evaluation" includes different models or data collection strategies to gather information at different stages in the life of a project. A major goal of this chapter is to help project directors and principal investigators understand what these are and how to use them.

As we undertake this discussion, it is important to recognize that within NSF there are two basic levels of evaluation: program evaluation and project evaluation. While this Handbook is directed at the latter, it is important to understand what is meant by both. Let us start by defining terms and showing how they relate.

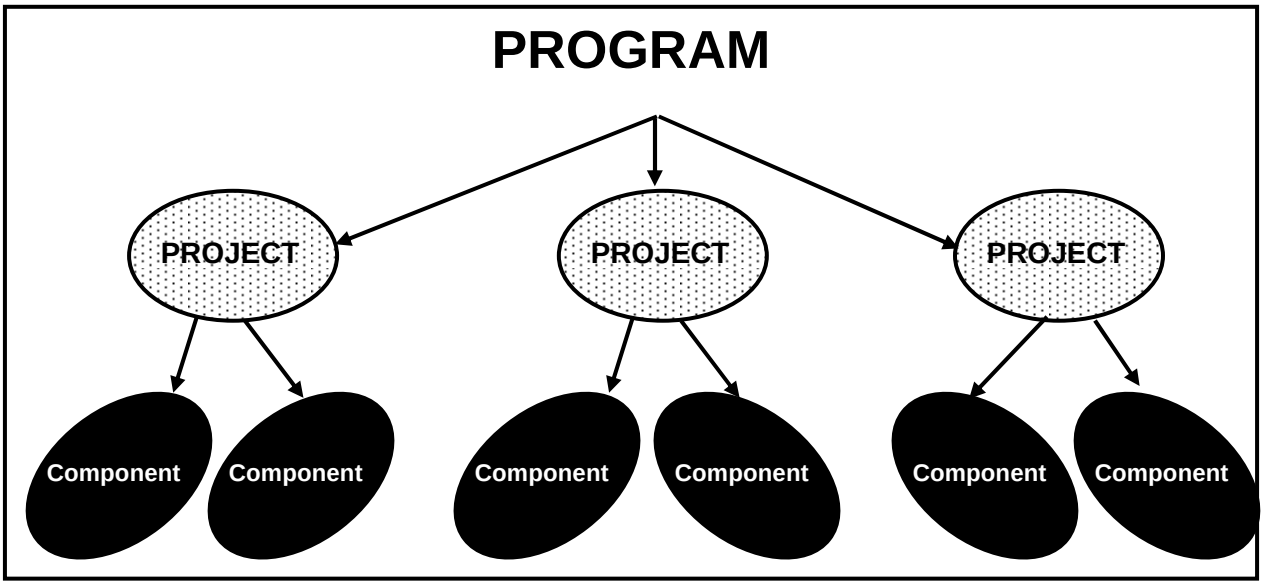
A **program** is a coordinated approach to exploring a specific area related to NSF's mission of strengthening science, mathematics, and technology. A **project** is a particular investigative or developmental activity funded by that program. NSF initiates a program on the assumption that an agency goal (such as increasing the strength and diversity of the scientific workforce) can be attained by certain educational activities and strategies (for example, providing supports to selected groups of undergraduate students interested in science or mathematics). The Foundation then funds a series of discrete projects to explore the utility of these activities and strategies in specific situations. Thus, a program consists of a collection of projects that seek to meet a defined set of goals and objectives.

Now let us turn to the terms "program evaluation" and "project evaluation." A **program evaluation** determines the value of this collection of projects. It looks across projects, examining the utility of the activities and strategies employed. Frequently, a full-blown program evaluation may be deferred until the program is well underway, but selected data on interim progress are collected on an annual basis. **Project evaluation**, in contrast, focuses on an individual project funded under the umbrella of the program. The evaluation provides information to improve the project as it develops and progresses. Information is collected to help determine whether the project is proceeding as planned and whether it is meeting its stated program goals and project objectives according to the proposed timeline. Ideally, the evaluation design is part of the project proposal, baseline data are collected prior to project initiation, and new data collection begins soon after the project is funded. Data are examined on an ongoing basis to determine if current operations are satisfactory or if some modifications might be needed.

Where a project consists of multiple components, evaluations might also include examination of specific components, as shown in Exhibit 2. A component of a project may be a specific teacher training approach, a

classroom practice, or a governance strategy. An evaluation of a component frequently examines the extent to which its goals have been met (these goals are a subset of the overall project goals), and seeks to clarify the extent to which the component contributes to the success or failure of the overall project.

**Exhibit 2.—Levels of evaluation**

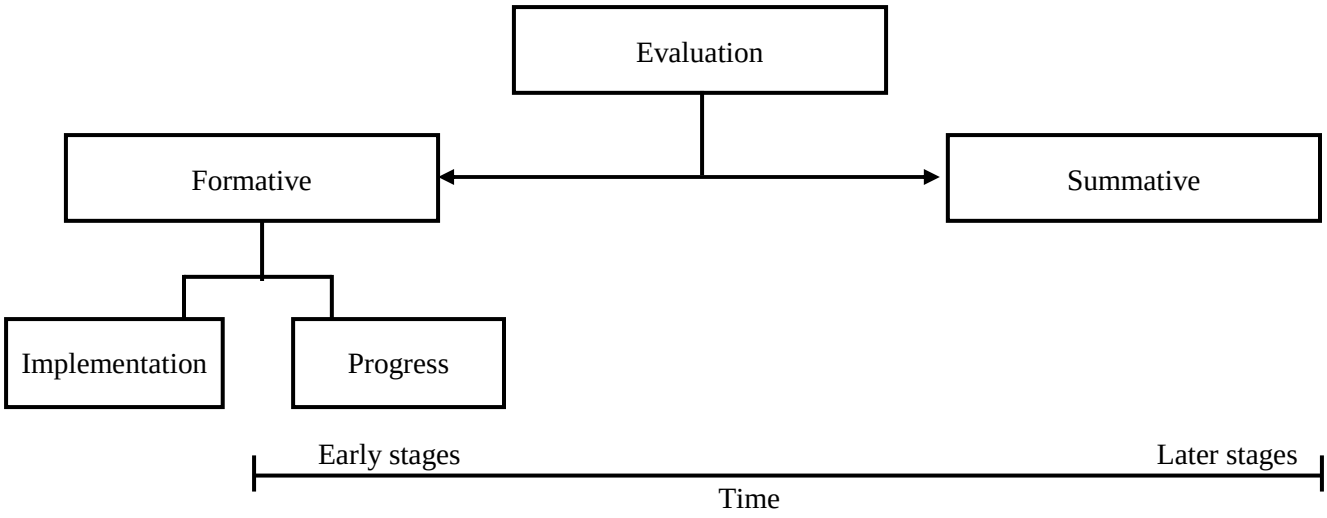



The information in this Handbook has been developed primarily for the use of project directors and principal investigators, although project evaluators may also find it useful. Our aim is to provide tools that will help those responsible for the examination of individual projects gain the most from their evaluation efforts. Clearly, however, these activities will also benefit program studies and the work of the Foundation in general. The better the information is about each of NSF's projects, the more we can all learn.

## The Different Kinds of Evaluation

Educators typically talk about two purposes for evaluation—formative evaluation and summative evaluation. The purpose of a formative evaluation is to provide information for project improvement. The purpose of a summative evaluation is to assess the quality and impact of a fully implemented project (see Exhibit 3).

**Exhibit 3.—Types of evaluation**




**Formative Evaluation**

Formative evaluation begins during project development and continues in some form throughout the life of the project. Its intent is to assess ongoing project activities and provide information to monitor and improve the project. According to evaluation theorist Bob Stake,

*A formative evaluation assesses ongoing project activities.*

> "When the cook tastes the soup, that's formative;
> When the guests taste the soup, that's summative."

Formative evaluation has two components: implementation evaluation and progress evaluation.

**Implementation Evaluation.** The purpose of implementation evaluation is to assess whether the project is being conducted as planned. This type of evaluation, sometimes called "process evaluation," typically occurs several times during the life of the grant or contract at least in multi-year projects. The underlying principle is that before you can evaluate the outcomes or impact of a project, you must examine how it is operating, whether it is operating according to the proposed plan or description, and whether some modification is needed.

*The purpose of implementation evaluation is to assess whether the project is being conducted as planned.*

In addition to assessing fidelity, implementation evaluation serves the purpose of describing and documenting the activities a project undertakes. This descriptive phase may be especially important in NSF programs that have a research and development (R&D) emphasis. In such programs, wide latitude is given to projects in what they are to do, as long as their plan is research-based and aligned to program goals. In such projects describing what is being done and, in combination with

progress evaluation (described below), identifying the strengths and weaknesses of different strategies becomes a critical first step.

A series of implementation questions guides an implementation evaluation. For example, NSF's Louis Stokes Alliances for Minority Participation (LSAMP) is aimed at increasing the quality and quantity of students successfully completing STEM baccalaureate degree programs, and increasing the number of students interested in, academically qualified for, and matriculated into programs of graduate study. LSAMP supports sustained and comprehensive approaches that facilitate achievement of the long-term goal of increasing the number of students who earn doctorates in STEM fields, particularly those from populations underrepresented in STEM fields. The program goals are accomplished through the formation of alliances.

Questions that might be posed for projects in the LSAMP program are as follows:

- Were appropriate students selected? Were students with deficits in precollege preparation included as well as ones with stronger records? Was the makeup of the participant group consistent with NSF's goal of developing a more diverse workforce?
- Were appropriate recruitment strategies used? Were students identified early enough in their undergraduate careers to provide the transitional supports needed?
- Were students given both academic and personal supports? To what extent were meaningful opportunities to conduct research provided?
- Was a solid project management plan developed and followed?

Sometimes the terms "implementation evaluation" and "monitoring evaluation" are confused. They are not the same. An implementation evaluation is an early check by the project staff, or the evaluator, to see if all essential elements are in place and operating. Monitoring is an external check. The monitor typically comes from the funding agency and is responsible for determining progress and compliance on a contract or grant for the project. Although the two differ, implementation evaluation, if effective, can facilitate the project's development and ensure that there are no unwelcome surprises during monitoring.

*The purpose of a progress evaluation is to assess progress in meeting the goals.*

**Progress Evaluation.** The purpose of a progress evaluation is to assess progress in meeting the project's ultimate goals. It involves collecting information to learn whether or not the benchmarks for progress were met and to point out any unexpected deviations. Progress evaluation collects information to determine what the impact of the activities and strategies is on participants, curricula, or institutions at various stages of the intervention. By measuring progress, the project can get an early indicator of whether or not project goals are likely to be achieved. If the data collected as part of

the progress evaluation fail to show expected changes, the information can be used to fine-tune the project. Data collected as part of a progress evaluation can also contribute to, or form the basis for, a summative evaluation conducted at some future date. In a progress evaluation of projects in the LSAMP program, the following questions can be addressed:

- Are the participants moving toward the anticipated goals of the project? Are they enhancing their academic skills? Are they gaining confidence in themselves as successful learners? Are they improving their understanding of the research process?
- Are the numbers of students reached by the projects increasing? How do changes in project participation relate to changes in the overall enrollments in mathematics, science, and technology areas at their institutions? Are students being retained in their programs at an increasing rate?
- Does student progress seem sufficient in light of the long-range goals of the program and project to increase the number of traditionally underrepresented students who receive degrees in science, mathematics, or technology?

Progress evaluation is useful throughout the life of the project, but it is most vital during the early stages when activities are piloted and their individual effectiveness or articulation with other project components is unknown.

**Summative Evaluation**

The purpose of summative evaluation is to assess a mature project's success in reaching its stated goals. Summative evaluation frequently addresses many of the same questions as a progress evaluation, but it takes place after the project has been established and the time frame posited for change has occurred. In addition, examining the extent to which the project has the potential to continue after the NSF funding cycle is completed—"sustainability"—is critical to NSF. Contributions to the broader knowledge base are also highly valued.

*The purpose of summative evaluation is to assess a mature project's success in reaching its stated goals.*

A summative evaluation of a project funded through the LSAMP program might address these basic questions:

- Are greater numbers of students from diverse backgrounds receiving bachelor's of science degrees and showing increased interest in scientific careers?
- Are there any impacts on the institutions of higher education the participants attend? Are there any changes in courses? Are there any impacts of the project on overall

*Summative evaluation collects information about outcomes and related processes, strategies, and activities that have led to them.*

course offering and support services offered by their institution(s)?

- Which components are the most effective? Which components are in need of improvement?
- Were the results worth the project's cost?
- Can the strategies be sustained?
- Is the project replicable and transportable?

Summative evaluation collects information about outcomes and related processes, strategies, and activities that have led to them. The evaluation is an appraisal of worth or merit. Usually this type of evaluation is needed for decision making about the future of the intervention. The decision alternatives may include the following: disseminate the intervention to other sites or agencies; continue funding; increase funding; continue on probationary status; modify and try again; or discontinue.

In most situations, especially high-stakes situations or those that are politically charged, it is important to have an external evaluator who is seen as knowledgeable, objective, and unbiased. Appendix A provides some tips for finding an evaluator. If that is not possible, it is better to have an internal evaluation than none at all. One compromise between the external and internal models is to conduct an internal evaluation and then hire an outside agent to both review the design and assess the validity of the findings and conclusions.

When conducting a summative evaluation, it is important to consider unanticipated outcomes. These are findings that emerge during data collection or data analyses that were never anticipated when the study was first designed. For example, consider an NSF project providing professional development activities for teacher leaders. An evaluation intended to assess the extent to which participants share their new knowledge and skills with their school-based colleagues might uncover a relationship between professional development and attrition from the teaching force. These results could suggest new requirements for participants or cautions to bear in mind.

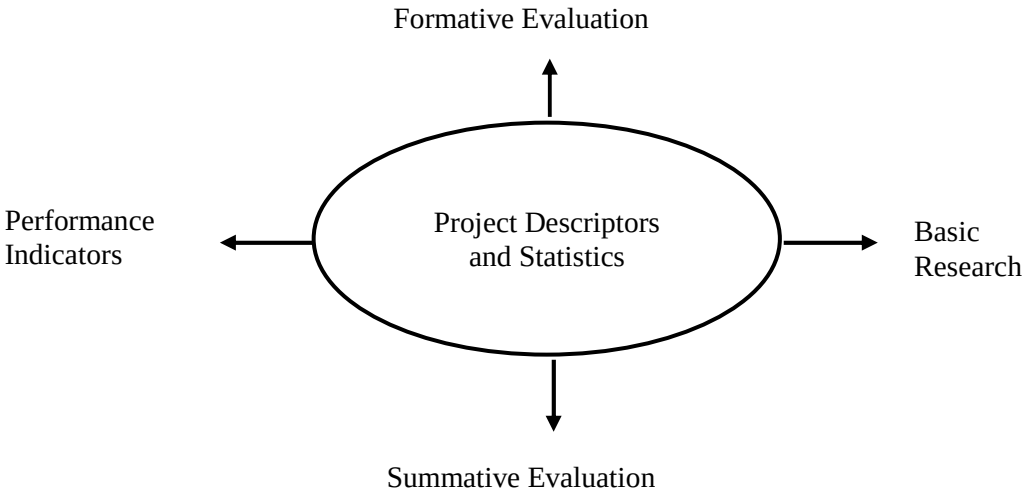
## Evaluation Compared to Other Types of Data Collection Activities

It is useful to understand how evaluation complements, but may differ from, other types of data collection activities that provide information on accountability for an NSF-funded project. Exhibit 4 shows various types of data collection activities, each of which provides somewhat different information and serves somewhat differing purposes. Included are performance indicators, formative evaluation, summative evaluation, and research studies.

*Evaluation complements but is different from other kinds of data collection activities.*

At the center of the effort is the project description, which provides general information about a project. These data are commonly used to monitor project activities (e.g., funding levels, total number of participants), to describe specific project components (e.g., duration of program activity, number of participants enrolled in each activity), and to identify the types of individuals receiving services. Descriptive information may be collected annually or even more frequently to provide a basic overview of a project and its accomplishments. Obtaining descriptive information usually is also part of each of the other data collection activities depicted. NSF has developed the FastLane system as one vehicle for collecting such statistics. FastLane allows for basic data to be collected across all programs in a consistent and systematic fashion. Some programs have added program-specific modules aimed at collecting tailored data elements or their projects.

**Exhibit 4.—Types of data collection activities**




Formative and summative evaluations are intended to gather information to answer a limited number of questions. Evaluations include descriptive information, but go well beyond that. Generally, formative and summative evaluations include in-depth data collection activities, are intended to support decision making, and range in cost, depending on the questions asked and project complexity.

Performance indicators fall somewhere between general program statistics and formative/summative evaluations. A performance indicator system is a collection of statistics that can be used to monitor the ongoing status of a program against a set of targets and metrics. Performance indicators play a critical role in the GPRA and PART activities described in the previous chapter. Going beyond project description, performance indicators begin to provide information that can be measured against a set of goals and objectives. Indicator systems are typically used to focus policymakers, educators, and the public on (1)

key aspects of how an educational program is operating, (2) whether progress is being made, and (3) where there are problems (Blank, 1993). Because performance indicators focus on tangible results, they often go beyond traditional reviews of project expenditures and activity levels. In fact, the term "performance" underscores the underlying purpose of indicator systems, i.e., to examine accomplishments of the projects in a program and measure progress toward specific goals. Performance indicators provide a snapshot of accomplishments in selected areas; however, in contrast to evaluations, the information is limited and is unlikely to provide an explanation of why a project may have succeeded or failed.

Research studies include descriptive information and provide targeted in-depth exploration of issues, but differ along other dimensions. Instead of being intended for decision making, research efforts typically are designed to broaden our understanding. They frequently explore conceptual models and alternative explanations for observed relationships.

## Using Evaluation Information

Earlier we defined evaluation as "systematic investigation of the worth or merit of an object." Why would someone want to assess worth or merit? There may be many reasons, but an important one is to make changes or improvements in the status quo. An evaluation document that sits on a shelf may provide proof that an activity occurred—thus meeting grant or contract requirements—but if "shelving" is all that results from an evaluation, it hardly seems worth the time and effort.

Patton (2008) argues for an approach to evaluation called "utilization-focused evaluation," the premise of which is "that evaluations should be judged by their utility and actual use" (p. 20). He makes a strong argument that simply generating evaluation findings is of relatively little importance compared to creating a context in which evaluation findings are actually used for decision making and improvement.

An important question that might be raised, therefore, as part of the overall evaluation task is whether or not the stakeholders of a particular project actually used the information in some way and to what end. Regarding utilization, many different stakeholders can be considered. The funder and project director are important, but depending on the project and its dissemination, there may be impacts at the grassroots—impacts among the doers—far sooner than among the managers.

As you develop your evaluations plans, you may want to consider including a component that looks more closely at the use of evaluation findings, either in the short term or after enough time has passed that change could reasonably be expected.

## Summary

The goals of evaluation are twofold: first, to provide information for project improvement and second, to determine the worth or merit of some procedure, project, process, or product. Well-designed evaluations also provide information that can help explain the findings that are observed and make a broader contribution to the knowledge base in the field. Increasingly, scientists, mathematicians, engineers, and educators are faced with the challenges of evaluating their innovations and determining whether progress is being made or stated goals have, in fact, been reached. Both common sense and accepted professional practice would suggest a systematic approach to these evaluation challenges. The role that evaluation may play will vary depending on the timing, the specific questions to be addressed, and the resources available. It is best to think of evaluation not as an event, but as a process. The goal should be to provide an ongoing source of information that can aid decision making at various steps along the way.

## References

Blank, R. (1993). Developing a System of Education Indicators: Selecting, Implementing, and Reporting Indicators. *Educational Evaluation and Policy Analysis*, 15 (1, Spring): 65-80.

Patton, M.Q. (2008). *Utilization-Focused Evaluation.* 4th Ed. Thousand Oaks, CA: Sage.

# Chapter 3 THE EVALUATION PROCESS—GETTING STARTED

In the preceding chapter, we outlined the types of evaluations that should be considered for projects funded through NSF's programs. In this chapter, we talk further about how to carry out an evaluation, expanding on the steps in evaluation design and development. Our aim is to provide an orientation to some of the basic language of evaluation, as well as to share some hints about technical, practical, and political issues that should be kept in mind when conducting project evaluations.

Whether they are summative or formative, evaluations can be thought of as having six phases:

- Development of a conceptual model of the program and identification of key evaluation points
- Development of evaluation questions and definition of measurable outcomes
- Development of an evaluation design
- Collection of data
- Analysis of data
- Provision of information to interested audiences

Getting started right can have a major impact on the progress and utility of the evaluation. However, all six phases are critical to providing a high-quality product. If the information gathered is not perceived as valuable or useful (the wrong questions were asked), or the information is not seen to be credible or convincing (the wrong techniques were used), or the report is presented too late or is not understandable (the teachable moment is past), then the evaluation will not serve its intended purpose.

*Getting started right can have a major impact on the progress and utility of the evaluation all along the way.*

In the sections below, we provide an overview of the first three phases, which lay the groundwork for the evaluation activities that will be undertaken. The remaining three phases are discussed in Chapter 4.

## Develop a Conceptual Model of the Project and Identify Key Evaluation Points

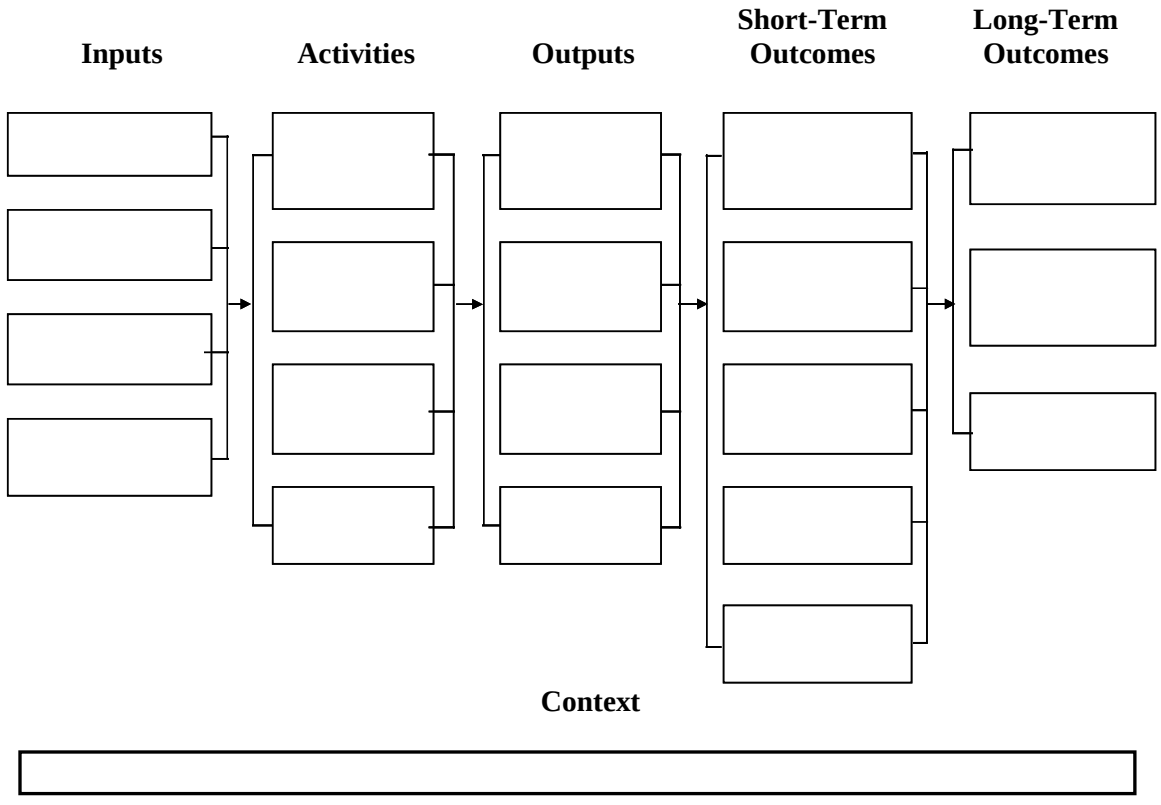
Every proposed evaluation should start with a conceptual model to which the design is applied. Conceptual models draw on both the research

literature and what might be called "craft knowledge"—experience-based understanding and hypotheses. This conceptual model can be used both to make sure that a common understanding about the project's structure, connections, and expected outcomes exists, and to assist in focusing the evaluation design on the most critical program elements.

Exhibit 5 presents the shell for a particular kind of conceptual model, a "logic model."[1] The model describes the pieces of the project and expected connections among them. A typical model has five categories of project elements that are connected by directional arrows. These elements are:

- Inputs
- Activities
- Outputs
- Short-Term and Long-Term Outcomes
- Context

**Exhibit 5.—Logic model**




[1] There are several different ways to show a logic model. The model presented here is one that was developed by the Kelly Foundation and has been useful to the author.

**Inputs** are the various funding sources and resource streams that provide support to the project. NSF funding is an input; in-kind contributions would also be an input.

**Activities** are the services, materials, and actions that characterize the project's thrusts. Developing a new preservice course would be an activity as would the provision of professional development.

**Outputs** are the products of these activities or a count that describes the activity. An output would be the number of hours of professional development offered.

**Outcomes** are the changes that occur as a result of the activities; they may be short or longer term. The acquisition of new content or pedagogical skills is an outcome.

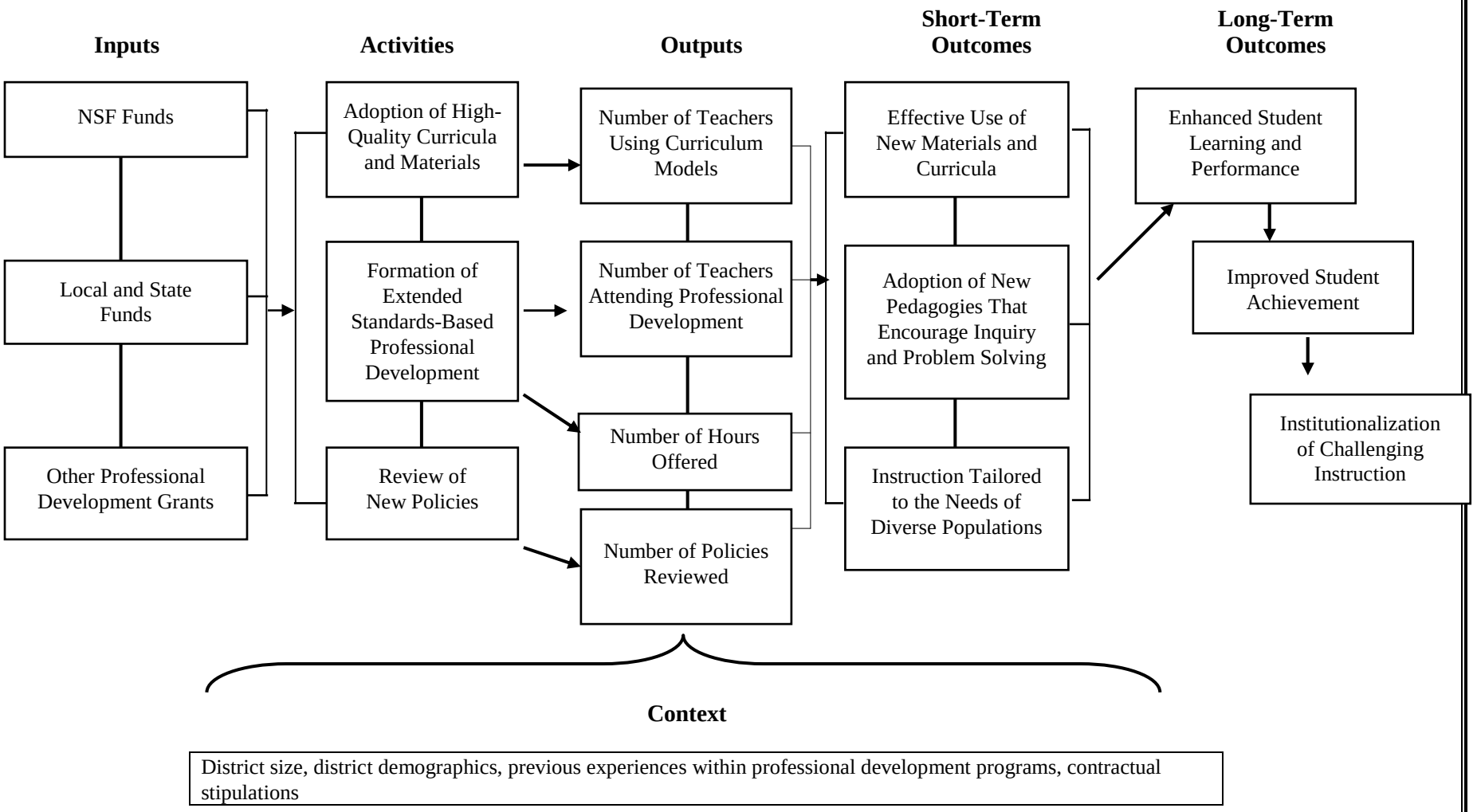
**Context** describes the specific features of a project that may affect its implementation and ultimate generalizability. Some examples of contextual variable would be demographics, state mandates, institutional policies, and economic conditions.

A logic model identifies these program elements and shows expected connections among them. Logic models are closely linked to approaches to evaluation that stress the importance of having a theory of change that underlies a project (Frechtling, 2007). This theory can be based on empirical research or practical experience. The purpose of the evaluation is to gather data than can test—affirm or reject—the proposed theory of change. Logic models make explicit the theory of change and thus help to guide the evaluator in selecting the questions to address and the linkages that need to be explored. They also contribute to decisions regarding appropriate methodologies to use. A logic model is a dynamic tool; as projects are modified in critical ways that reflect modifications in the underlying theory, it is important to revise and update the logic model.

The visual of the logic model above shows a process that flows left to right from inputs to long-term outcomes. This flow and unidirectionality is misleading for a number of reasons. First, in developing a model for your project, it may be useful to reverse this flow. That is, project teams frequently find it more useful to "work backwards," starting from the long-term outcome desired and then determining critical conditions or events that will need to be established before these outcomes might be expected to occur. Second, using logic models frequently results in feedback loops that change the model from one that is unidirectional to one that is multidirectional. For example, early findings often result in changes in activities that must be reexamined for both implementation and outcomes as a project evolves.

Exhibits 6 and 7 show logic models for two NSF programs: Local Systemic Change (LSC) and ADVANCE. The LSC program supported intensive professional development projects that combined inservice

**Exhibit 6.—Conceptual model for Local Systemic Change (LSC) Initiatives**



support with the use of high-quality, standards-based curricula. Local school districts involved in the program were expected to engage teachers in 120 hours of professional development support over the funding period. The program addressed both mathematics and science across the K–12 grade spectrum.

Under inputs, we have listed three streams of funding:

- NSF funds
- Local and state funds
- Other professional development grants

Contextual factors include:

- District size and demographics
- Previous experiences with professional development programs
- Contractual stipulations

For "activities," we have highlighted:

- Adoption of high-quality curricula and materials
- Formation of extended standards-based professional development
- Review of new policies

Some outputs of interest are:

- Number of teachers using curriculum models
- Number of teachers attending professional development
- Number of hours of professional development offered
- Number of policies reviewed

The short-term outcomes are linked to, and flow from, the overall goals of the LSCs. Thus, we would look for:

- Effective use of new materials and curricula
- Adoption of new pedagogies that encourage inquiry and problem solving

- Instruction tailored to the individual needs of students from diverse populations

Finally, over time, the LSCs should result in:

- Enhanced student learning and performance
- Higher scores on assessments of student achievement
- Institutional change

The ADVANCE IT Logic Model, developed for the evaluation of the Institutional Transformation component of the ADVANCE program (Berkowitz et al. 2009) is somewhat more complex. Further, because of the expectation for broader based effects, this model also includes "impacts" or systemic changes in a field or endeavor.

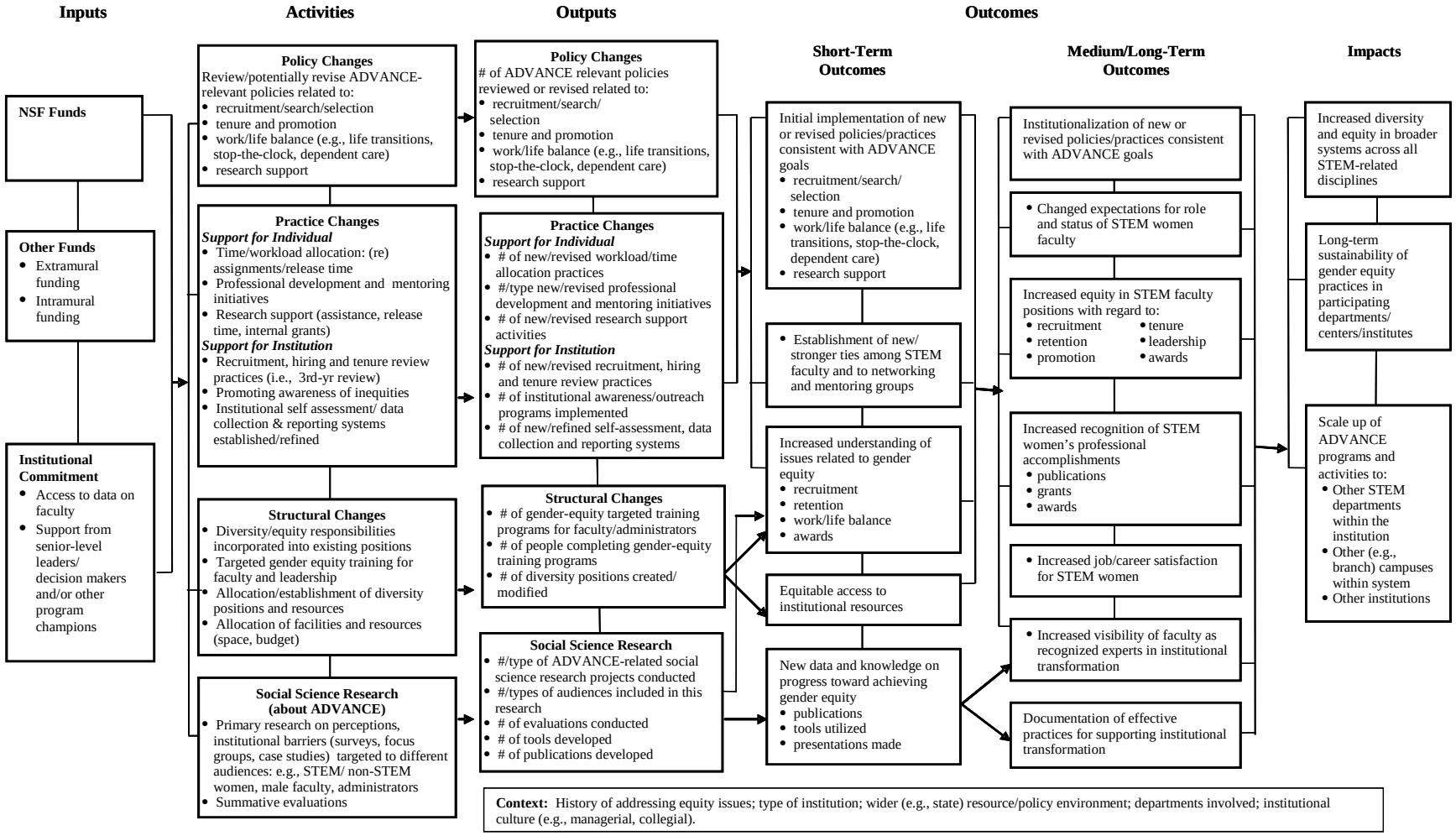
Under inputs, we have financial and non-financial resources

- Financial resources include both NSF monies for ADVANCE-related activities and other funding from both internal and external sources that help to support similar or complementary efforts to achieve gender equity in the participating IHEs.
- Non-financial resources include the level and nature of institutional commitment to the IT project, including support from senior-level decision makers, existence of a program champion or champions and provision of ready access to data on faculty, including both men and STEM women faculty.

Contextual factors include:

- History and experience addressing equity issues, especially gender equity issues
- Type of institution
- Departments/disciplines/centers and/or institutes participating in ADVANCE
- Institutional culture
- Wider resource and policy environments within which the IHEs are functioning

## Exhibit 7. Logic model for ADVANCE IT Program



For "activities," we have highlighted:

- Policy Changes that involve activities undertaken to review, and potentially revise, STEM faculty policies; formation of extended standards-based professional development.

- Practice Changes including changes aimed at individual STEM women faculty members and changes in institutional level practices.

- Structural Changes that involve activities aimed at making changes in the organization of the participating IHE that promote greater institutional attention and commitment to addressing gender inequities.

- Social Science Research Activities directed at increasing knowledge of institutional barriers to and facilitators of advancement of STEM women faculty, as well as exemplary practices for overcoming these barriers.

Some outputs of interest are:

- Number of relevant policies reviewed and/or revised in each of the four specified areas (recruitment/search/selection; tenure and promotion; work/life balance; and research support).

- Number of new or revised practices related to equitable allocation or reallocation of individual faculty members' workload.

- Number of new or revised recruitment, hiring, and tenure review practices developed in the participating IHE.

- Number and types of ADVANCE-related social science research activities carried out.

- Number of related tools and/or publications developed.

The short-term outcomes are linked to, and flow from, the overall goals of the ADVANCE. Thus, we would look for

- Initial implementation of new or revised policies and practices consistent with ADVANCE goals in any or all of the relevant areas: recruitment, search and/or selection, tenure/promotion, work-life balance, and research support.

- STEM faculty establishing stronger ties with one another through networking and developing mentoring relationships to individuals and groups both at the home institution and in the wider field of research.

- Achieving greater institutional recognition and understanding of issues related to gender equity in key areas, as well as equitable access to institutional resources.

- New data and knowledge on how to make progress toward achieving gender equity in IHEs.

In the long term, we should expect to see at participating institutions

- Institutionalization of new or revised gender-equity promoting policies and practices.

- Changed expectations with respect to the role and status of STEM women faculty.

- Increased equity in STEM faculty positions.

- Increased acknowledgment and recognition of STEM women faculty's professional accomplishments as well as STEM women's increased satisfaction with their jobs and academic careers.

- Increased visibility of faculty at the IHEs who are recognized experts in institutional transformation.

- Documentation of effective practices for supporting institutional transformation.

Finally, the logic model shows impacts.

- The "ultimate" global impact, shown at the top of the model, is increased diversity and equity in broader systems across all STEM and STEM-related disciplines.

- Other systemic impacts are conceived at two levels.

  - The first of these is the long-term sustainability of all the gender equity promoting policies and practices—and so their attendant outcomes—in the participating departments, centers, and institutes.

  - The other such set of impacts has to do with scale-up of ADVANCE IT programs and activities to other departments within the participating institution, other campuses within the system, and other institutions of higher education. In this way the examples set by the participating institutions, and the lessons learned from them, can be disseminated through the broader system.

Once this logic model is developed and connections are established, the next step is to clarify the timing for when the activities and impacts would be expected to emerge. This area should have been addressed during the project's planning phase, and determining expected time frames should be a revisiting of decisions rather than creation of a set of new considerations. However, either because some aspect was overlooked in the initial discussions or some conditions have changed, it is important to review the time schedule and make sure that the project is willing to be held accountable for the target dates.

Principal investigators and project directors may also find the logic model useful for project management. It provides a framework for monitoring the flow of work and checking whether required activities are being put in place as expected.

## Develop Evaluation Questions and Define Measurable Outcomes

The development of evaluation questions builds on the conceptual model and consists of several steps:

- Identifying key stakeholders and audiences
- Formulating potential evaluation questions of interest to the stakeholders and audiences
- Defining outcomes in measurable terms
- Prioritizing and eliminating questions

While it is obvious that NSF program managers and the directors of individual projects are key stakeholders in any project, it is important in developing the evaluation design to go beyond these individuals and consider other possible audiences and their needs for information. In all projects, multiple audiences exist. Such audiences include the participants, would-be participants, community members, NSF scientists, school administrators, parents, etc. Further, some of the audiences may themselves be composed of diverse groups. For example, most educational interventions address communities made up of families from different backgrounds with different belief structures. Some are committed to the status quo; others may be strong advocates for change.

*It is important to identify stakeholders early in the design phase.*

In developing an evaluation, it is important to identify stakeholders early in the design phase and draw upon their knowledge as the project is shaped. A strong stakeholder group can be useful at various points in the project—shaping the questions addressed, identifying credible sources of evidence, and reviewing findings and assisting in their interpretation. Getting stakeholders involved early on

may also increase the credibility of the evaluation and the likelihood that the information will be used.

Although, in most cases, key stakeholders will share a number of information needs (in a professional development project the impacts on teaching quality will be of interest to all), there may be audience-specific questions that also need to be considered. For example, while exposure to the new technologies in an NSF lab may provide teachers with important new skills, administrators may be concerned not only with how the introduction of these skills may impact the existing curriculum, but also in the long-term resource and support implications for applying the new techniques. Depending on the situation and the political context in which a project is being carried out, a judicious mix of cross-cutting and audience-specific issues may need to be included. Exhibit 8 presents a shell for organizing your approach to identifying stakeholders and their specific needs or interests.

The process of identifying potential information needs usually results in many more questions than can be addressed in a single evaluation effort. This comprehensive look at potential questions, however, makes all of the possibilities explicit to the planners of the evaluation and allows them to make an informed choice among evaluation questions. Each potential question should be considered for inclusion on the basis of the following criteria:

- The contribution of the information to the goals of NSF and the projects' local stakeholders
- Who would use the information
- Whether the answer to the question would provide information that is not now available
- Whether the information is important to a major group or several stakeholders
- Whether the information would be of continuing interest
- How the question can be translated into measurable terms
- How it would be possible to obtain the information, given financial and human resources

**Exhibit 8.—Identifying key stakeholders**

| List the audiences for your evaluation | Identify persons/spokespersons for each audience | Describe the particular values, interests, expectations, etc., that may play a key role as criteria in the analysis and interpretation stage of your evaluation |
|---|---|---|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

These latter two points require some additional explanation. First is the question of measurability. There are some evaluation questions that, while clearly important, are very challenging to address because of the difficulty of translating an important general goal into something that can be measured in a reliable and valid way. For example, one of the goals of a summer research experience for teachers might be generally stated "to increase the extent to which teachers use standards-based instruction in their science teaching." To determine whether or not this goal is met, the evaluation team would have to define an indicator or indicators of standards-based instruction, establish a goal for movement on the part of the teachers, and then set interim benchmarks for measuring success. A variety of possible articulations exist. One could talk about the percentage of teachers moving through various levels of proficiency in standards-based instruction (once those levels were established); or the outcome could be measured in terms of the percentage of time devoted to different practices; or understanding, rather than actual practice, could be examined. Each approach probably has strengths and weaknesses. The critical task is to determine a shared definition of what is meant and what will be accepted as credible evidence of project success.

Exhibits 9a and 9b illustrate the steps taken to translate a general goal into a measurable objective. Using the LSAMP example previously discussed, we get the following (see Exhibit 9b).

P**urpose of project.** As stated in the previous chapter, LSAMP aims to increase the quality and quantity of students successfully completing STEM baccalaureate degree programs, and increase the number of students interested in, academically qualified for, and matriculated into programs of graduate study.

**State a general goal.** To increase the quantity of students successfully completing a STEM baccalaureate degree.

**Define an objective.** To increase the percentage of students initially declaring a STEM major that actually graduate with a STEM major.

**Break the objective down further.** To increase the percentage of students from underrepresented minority groups who declare a STEM major that actually graduate with a STEM major.

**Make sure the objective is measurable or restate it.** This objective is measurable in terms of declared major and observed graduation status.

**State the criteria for success.** Over the five years of the project, the percentage of declared STEM majors who earn baccalaureates in the STEM field will increase by 50 percent.

A particular challenge in developing measurable objectives is determining the criteria for success, that is, deciding how much change is enough to declare the result important or valuable. The classical approach to this question is to look for changes that are statistically significant, i.e., typically defined as unlikely to occur by chance in more than one to five percent of the observations. While this criterion is important, statistical significance may not be the only or even the best standard to use. If samples are large enough, a very small change can be statistically significant. When samples are very small, achieving statistical significance may be close to impossible.

What are some ways of addressing this problem of the importance or meaningfulness of change? First, for large samples, "effect size" is frequently used as a standard against which to measure the *importance* of an outcome. Using this approach, the amount of change is measured against the standard deviation, and only those significant outcomes that result in a change of a certain amount are considered meaningful. Generally, effect sizes of .25 or more are considered important by the researcher (Bloom, 2005) and the *What Works Clearinghouse*. Second,

**Exhibit 9a.—Goal and objective writing worksheet**

# GOAL AND OBJECTIVE WORKSHEET

1. Briefly describe the purpose of the project.

2. State the above in terms of a general goal.

3. State an objective to be evaluated as clearly as you can.

4. Can this objective be broken down further? Break it down to the smallest unit. It must be clear what specifically you hope to see documented or changed.

5. Is this objective measurable (can indicators and standards be developed for it)? If not, restate it.

6. Using the indicator described above, define the criteria for success.

7. Once you have completed the above steps, go back to #3 and write the next objective. Continue with steps 4, 5, and 6.

**Exhibit 9b.—Sample goal and objective writing worksheet for an LSAMP goal**

# GOAL AND OBJECTIVE WORKSHEET

1. Briefly describe the purpose of the project.

   LSAMP aims to increase the quality and quantity of students successfully completing STEM baccalaureate degree programs, and increase the number of students interested in, academically qualified for, and matriculated into programs of graduate study.

2. State the above in terms of a general goal.

   Increase the quantity of students successfully completing a STEM baccalaureate degree.

3. State an objective to be evaluated as clearly as you can.

   Increase the percentage of students initially declaring a STEM major that actually graduate with a STEM major.

4. Can this objective be broken down further? Break it down to the smallest unit. It must be clear what specifically you hope to see documented or changed.

   Increase the percentage of students from underrepresented minority groups who declare a STEM major that actually graduate with a STEM major.

5. Is this objective measurable (can indicators and standards be developed for it)? If not, restate it.

   Measurable in terms of declared major and observed graduation status.

6. Using the indicator described above, define the criteria for success.

   Over the five years of the project, the percentage of declared STEM majors who earn baccalaureates in the STEM field will increase by 50 percent.

it may be possible to use previous history as a way of determining the importance of a statistically significant result. The history can provide a realistic baseline against which the difference made by a project can be assessed.

Third, with or without establishing statistical significance, expert judgment may be called on as a resource. This is a place where stakeholder groups can again make a contribution. Using this approach, standards are developed after consultation with differing stakeholder groups to determine the amount of change each would need to see to find the evidence of impact convincing.

There is also the issue of feasibility of carrying out the measurement, given available resources. Three kinds of resources need to be considered: time, money, and staff capability. The presence or absence of any of these strongly influences whether or not a particular question can be addressed in any given evaluation. Specifically, there are some questions that may require specialized expertise, extended time, or a large investment of resources. In some cases, access to these resources may not be readily available. For example, it might be considered useful conceptually to measure the impact of a student's research experience in terms of the scientific merit of a project or presentation that the student completes before the end of a summer program. However, unless the evaluation team includes individuals with expertise in the particular content area in which the student has worked, or can identify consultants with the expertise, assessing scientific merit may be too much of a stretch. Under these circumstances, it is best to eliminate the question or to substitute a reasonable proxy, if one can be identified. In other cases, the evaluation technique of choice may be too costly. For example, classroom observations are valuable if the question of interest is "How has the XYZ Math and Science Partnership project affected classroom practices?" But observations are both time-consuming and expensive. If sufficient funds are not available to carry out observations, it may be necessary to reduce the sample size or use another data collection technique, such as a survey. A general guideline is to allocate five to 10 percent of project cost for the evaluation.

*A general guideline is to allocate five to 10 percent of project cost for the evaluation.*

## Develop an Evaluation Design

The next step is developing an evaluation design. Developing the design includes:

- Determining what type of design is required to answer the questions posed
- Selecting a methodological approach and data collection instruments
- Selecting a comparison group

- Sampling
- Determining timing, sequencing, and frequency of data collection

**Determining the Type of Design Required to Answer the Questions Posed**

There are many different types of evaluation designs that can be used. The selection among alternatives is not just a matter of evaluator preference but is strongly affected by the type of question that the project is trying to address. Shavelson and Towne (2002) identify three types of questions. Although they are posed as research questions, they apply equally well to evaluation questions. The three types of questions are:

- What is happening?
- Is there a systematic effect?
- Why or how is it happening?

Depending on which one or ones of these questions your evaluation is to address, the requirements for design will differ substantially. For example, if your question concerns what is happening, the focus of your evaluation work will be formative and targeted to a comprehensive description of the activities being implemented or materials developed. If, on the other hand, the question of principal interest concerns systematic effect, greater attention will need to be paid to employing a design that goes beyond the chronicling of project activities and their impacts and provides ways of establishing causal attribution and ruling out competing hypotheses. (This issue is discussed further in the section Selecting a Comparison Group.) Further, most investigators want to not only confirm an impact but also understand which of the potential features of the project are most influential in leading to that impact. This question is typically addressed through additional types of analyses.

**Selecting a Methodological Approach**

In developing the design, two general methodological approaches—quantitative and qualitative—frequently have been considered as alternatives. Aside from the obvious distinction between numbers (quantitative) and words (qualitative), the conventional wisdom among evaluators is that quantitative and qualitative methods have different strengths, weaknesses, and requirements that will affect evaluators' decisions about which are best suited for their purposes.

In Chapter 5, we review the debate between the protagonists of each of the methods and make a case for what we call a "mixed-methods" design. This is an approach that combines techniques traditionally

labeled "quantitative" with those traditionally labeled "qualitative" to develop a full picture of why a project may or may not be having hoped-for results and to document outcomes. A number of factors need to be considered in reaching a decision regarding the methodologies that will be used. These include the questions being addressed, the time available, the skills of the existing or potential evaluators, and the type of data that will be seen as credible by stakeholders and critical audiences.

**Selecting a Comparison Group**

In project evaluation, especially summative evaluation, the objective is to determine whether or not a set of experiences or interventions results in a set of expected outcomes. The task is not only to show that the outcomes occurred, but to make the case that the outcomes can be attributed to the intervention and not to some other factors. This is Shavelson and Towne's second type of research question.

In classical evaluation design, this problem of attribution is addressed by randomly assigning the potential pool of participants into treatment and control or comparison groups. In the ideal world, NSF project evaluators would adopt this same approach so that competing hypotheses can be ruled out and external validity increased. A detailed discussion of factors to keep in mind in designing true experiments that meet the highest quality standards can be found in the *What Works Clearinghouse Study Review Standards* (2003).

Evaluators face two basic challenges: deciding when it is appropriate to conduct an experimental study and obtaining the sample needed to carry one out. Experimental studies are difficult to do and require considerable resources to carry out well (Rossi, Lipsey, and Freeman, 2003). Projects that are at the proof of concept stage or are beginning to explore the utility of an approach are probably not the best candidates for experimental work—in all likelihood these are "what is happening" studies and require a focus on describing the intervention. Experimental studies typically are conducted when there is some evidence of success from preliminary work. Once such evidence is found, the next step is to test the intervention or activity under more rigorously controlled situations.

*In designing an evaluation it is important to address, rather than ignore, the attribution question.*

Having determined that an experimental study is appropriate, the next challenge is to determine a strategy for creating comparable treatment and comparison groups. There is no perfect way to do so, but if claims of effectiveness are to be made, every effort also must be made to create, or come as close as possible to creating, an evaluation design that meets what might be considered the "gold standard." Sometimes that can be done by providing attractive incentives for agreeing to be randomly assigned to either a treatment or a control group. Another strategy is to draw a comparison group from a waiting list (when one exists) and compare those who participated with those who also self-selected to participate but applied too late. Relatedly, when there are a sufficient

number of applicants, those who apply could be randomly assigned to two groups: one that receives the treatment initially, and one that serves as a comparison group for a given length of time and then is allowed to participate.

If a true experimental design cannot be constructed, a quasi-experimental design, in which a matched (but not randomly assigned) comparison group is included, is a good fallback position. In other cases, it may be possible to use historical data as a benchmark against which to measure change, such as comparing a school's previous test score history to test scores after some experience or intervention has taken place. If the historical approach is adopted, it is important to rule out other events occurring over time that might also account for any changes noted. In dealing with student outcomes, it is also important to make sure that the sample of students is sufficiently large to rule out differences associated with different cohorts of students. To avoid what might be called a "crop effect," it is useful to compare average outcomes over several cohorts before the intervention with average outcomes for multiple cohorts after the intervention.

A third alternative is to look for relationships between levels of implementation of some program and the outcome variable(s) of interest (Horizon and Westat, 2001). To some extent, a set of internal comparison groups is created by drawing on actual implementation data or a surrogate, such as years in the program or level of treatment. For example, in a teacher enhancement project where teachers received different amounts of professional development, subgroups could be created (derived from teacher surveys and/or classroom observation) to categorize classrooms into high, medium, and low implementation status (assuming amount of professional development received is not correlated with some factor that might confound the interpretation of results). With this approach, the outcome of interest would be differences among the project subgroups. It is assumed in this design that there is generally a linear relationship between program exposure or implementation and change along some outcome dimension. The evaluation thus examines the extent to which differences in exposure or implementation relate to changes in outcomes. Here, too, it is important to examine the extent to which the groups are comparable on other variables that might relate to the outcome of interest to rule them out as competing explanations for differences that might be found.

Finally, checking the actual trajectory of change against the conceptual trajectory, as envisioned in the logic model, often provides support for the likelihood that impacts were, in fact, attributable to project activities. Confirmation does not, however, translate into proof of causality.

Evaluators should strive to use true experiments whenever appropriate and possible, but are cautioned not to abandon project evaluation if the requirements of a true experiment cannot be met. While the alternative methods do not provide the strong evidence that is obtained through a true experiment, they do add value and contribute to the knowledge base.

**Sampling**

Except in rare cases when a project is very small and affects only a few participants and staff members, it is necessary to deal with a subset of sites and/or informants for budgetary and managerial reasons. Sampling thus becomes an issue in the development of an evaluation design. And the approach to sampling will frequently be influenced by the type of data collection method that has been selected.

The preferred sampling methods for quantitative studies are those that enable evaluators to make generalizations from the sample to the universe, e.g., all project participants, all sites, all parents. Random sampling is the appropriate method for this purpose. However, random sampling is not always possible.

The most common misconception about sampling is that large samples are the best way of obtaining accurate findings. While it is true that larger samples will reduce **sampling error** (the probability that if another sample of the same size were drawn, different results might be obtained), sampling error is the smallest of the three components of error that affect the soundness of sample designs. Two other errors—**sample bias** (primarily due to loss of sample units) and **response bias** (responses or observations that do not reflect "true" behavior, characteristics, or attitudes)—are much more likely to jeopardize validity of findings (Sudman, 1976). When planning allocation of resources, evaluators should give priority to procedures that will reduce sample bias and response bias, rather than to the selection of larger samples.

*When planning allocation of resources, evaluators should give priority to procedures that will reduce sample bias and response bias, rather than to the selection of larger samples.*

Let us talk a little more about sample and response bias. Sample bias occurs most often because of nonresponse (selected respondents or units are not available or refuse to participate, or some answers and observations are incomplete). Response bias occurs because questions are misunderstood or poorly formulated, or because respondents deliberately equivocate (for example, to protect the project being evaluated). In observations, the observer may misinterpret or miss what is happening. Exhibit 10 describes each type of bias and suggests some simple ways of minimizing it.

**Exhibit 10.—Three types of errors and their remedies**

| Type | Cause | Remedies |
|---|---|---|
| Sampling Error | Using a sample, not the entire population to be studied. | Larger samples, which reduce but do not eliminate sampling error. |
| Sample Bias | Some of those selected to participate did not do so or provided incomplete information. | Repeated attempts to reach nonrespondents; prompt and careful editing of completed instruments to obtain missing data; comparison of characteristics of nonrespondents with those of respondents to describe any suspected differences that may exist. |
| Response Bias | Responses do not reflect "true" opinions or behaviors because questions were misunderstood or respondents chose not to tell the truth. | Careful pretesting of instruments to revise misunderstood, leading, or threatening questions. No remedy exists for deliberate equivocation in self-administered questionnaires, but it can be spotted by careful editing. In personal interviews, this bias can be reduced by a skilled interviewer. |

Statistically valid generalizations are seldom a goal of qualitative evaluation; rather, the qualitative investigation is primarily interested in locating information-rich cases for study in depth. Purposeful sampling is therefore practiced, and it may take many forms. Instead of studying a random sample or a stratified sample of a project's participants, an evaluation may focus on the lowest achievers admitted to the program, or those who have never participated in a similar program, or participants from related particular regions. In selecting classrooms for observation of the implementation of an innovative practice, the evaluation may use deviant-case sampling, choosing one classroom where the innovation is reported as "most successfully" implemented and another where major problems are reported. Depending on the evaluation questions to be answered, many other sampling methods, including maximum variation sampling, critical case sampling, or even typical-case sampling, may be appropriate (Patton, 2001). The appropriate size of the sample may also differ when the different methodologies are adopted, with precision in numbers based on statistical considerations playing a much larger role for the quantitative approach.

In many evaluations, the design calls for studying a population at several points in time, e.g., students in the 9th grade and then again in the 12th grade. There are two ways to do this. In a longitudinal approach, data are collected from the same individuals at designated time intervals; in a cross-sectional approach, new samples are drawn for each successive data collection. While longitudinal designs that require collecting information from the same students or teachers at several points in time are best in most cases, they are often difficult and expensive to carry out both because students and teachers move and because linking individuals' responses over time is complicated. Furthermore, loss of respondents because of failure to locate or to obtain cooperation from some segments of the original sample is often a major problem. Depending on the nature of the evaluation and the size of the population studied, it may be possible to obtain good results with cross-sectional designs.

### Timing, Sequencing, and Frequency of Data Collection

*Project evaluations are stronger when data are collected in at least two points in time.*

The evaluation questions and the analysis plan largely determine when data should be collected and how often various data collections should be scheduled. In mixed-methods designs, when the findings of qualitative data collection affect the structuring of quantitative instruments (or vice versa), proper sequencing is crucial. As a general rule, project evaluations are stronger when data are collected at least two points in time: before an innovation is first introduced and after it has been in operation for a sizable period of time. Studies looking at program sustainability need at least one additional point of evidence: data on the program after it has been established and initial funding is completed.

All project directors find that both during the design phase, when plans are being crafted, and later, when fieldwork gets underway, some modifications and tradeoffs are necessary. Budget limitations, problems in accessing fieldwork sites and administrative records, and difficulties in recruiting staff with appropriate skills are among the recurring problems that should be anticipated as far ahead as possible during the design phase, but that also may require modifying the design at a later time.

What tradeoffs are least likely to impair the integrity and usefulness of an evaluation, if the evaluation plan as designed cannot be fully implemented? A good general rule for dealing with budget problems is to sacrifice the number of cases or the number of questions to be explored (this may mean ignoring the needs of some low-priority stakeholders), but to preserve the depth necessary to fully and rigorously address the issues targeted. If you are having problems gaining cooperation, you may need to transfer resources from services to incentives for participation. Sometimes a year of planning must be substituted for a year of implementation in complex studies.

Once decisions are reached regarding the actual aspects of your evaluation design, it is useful to summarize these decisions in a design matrix. Exhibits 11a and 11b present the shell for each matrix using projects from the Minority Research Fellowship Program (MRFP) as an illustrative example. This matrix is also very useful later on when it is time to write a final report (see Chapter 4).

**Exhibit 11a.—Matrix showing crosswalk of study foci and data collection activities**

| Study focus | Data collection activities | | | | |
|---|---|---|---|---|---|
| | Document review | Mail survey | Telephone interviews | Bibliometric measures | National data analysis |
| What did MRFP awardees do during their award period? In an extension if granted? | ✓ | ✓ | ✓ | | |
| Specifically, and as appropriate for postdoctoral scholars, to what extent have the individual research projects of the postdoctoral Fellows achieved their narrower and immediate scientific goals? To what extent is this reflected in the formal scientific record as publications and presentations? | ✓ | ✓ | ✓ | ✓ | |
| How if at all did MRFP awardees use their experience to shape their career direction and development? | ✓ | ✓ | ✓ | | |
| How do employment and activity patterns among MRFP awardees compare with patterns in national data on Ph.D. recipients who have been postdoctoral researchers? How does the NSF proposal and award history of MRFP awardees compare with that of other faculty members who received Ph.D.s in the fields and time period covered by the MRFP awardees? | | ✓ | ✓ | | ✓ |

**Exhibit 11b.—Crosswalk of study sample and data collections activities**

| Study sample | Data collection activities | | | | |
|---|---|---|---|---|---|
| | Document review | Mail survey | Telephone interviews | Bibliometric measures | National data analysis |
| All MRFP awardees (n=157) | ✓ | ✓ | | ✓ | ✓ |
| Sample of MRFP awardees (n=30) | | | ✓ | | |

## References


Berkowitz, S., Silverstein, G., Frechtling, J., Zhang, X., Lauman, B., Putman, H., & Segal, E. (2009) *Evaluation Plan: Quantitative Evaluation of the ADVANCE Program.*

Bloom, H. (2005). Randomizing groups to evaluate place-based programs. In *Learning More from Social Experiments: Evolving Analytical Approaches,* edited by H.S. Bloom, 115–172. New York: Russell Sage.

Department of Education. (2003). *What Works Clearinghouse Study Review Standards.*

Frechtling, J. (2007). *Logic Modeling Methods in Program Evaluation*. San Francisco, CA: Jossey-Bass.

Horizon and Westat. (2001). *Revised Handbook for Studying the Effects of the LSC on Students*. Rockville, MD: Westat.

Patton, M.Q. (2001). *Qualitative Evaluation and Research Methods*. 3rd Ed. Newbury Park, CA: Sage.

Rossi, P., Lipsey, M., and Freeman, H. (2004). *Evaluation: A Systemic Approach.* 7th Ed. Thousand Oaks, CA: Sage.

Shavelson, R., and Towne, L. (Eds.). (2002). *Scientific Research in Education*. Washington, DC: National Academy Press.

Sudman, S. (1976). *Applied Sampling*. New York: Academic Press.

Chapter 4

# THE EVALUATION PROCESS: CARRYING OUT THE STUDY AND REPORTING

In this section we discuss the steps to be undertaken after a design has been developed:

- Conducting the data collection
- Analyzing the data
- Reporting the findings
- Disseminating the information

## Conducting the Data Collection

Once the appropriate information-gathering techniques have been determined, the information must be gathered. Both technical and political issues need to be addressed.

- Obtain necessary clearances and permission.
- Consider the needs and sensitivities of the respondents.
- Make sure your data collectors are adequately trained and will operate in an objective, unbiased manner.
- Obtain data from as many members of your sample as possible.
- Cause as little disruption as possible to the ongoing effort.

*Many groups, especially school systems, have a set of established procedures for gaining clearance to collect data on students, teachers, or projects.*

First, before data are collected, the necessary clearances and permission must be obtained. Many school systems have a set of established procedures for gaining clearance to collect data on students, teachers, or projects. Issues may include identification of persons to receive/review a copy of the report, restrictions on when data can be collected, and procedures to safeguard the privacy of students or teachers. Parental permission is frequently a requirement for children, and informed consent may be required for adults. Universities have their own set of review requirements, with Institutional Review Board approval being required almost universally. It is important to find out what these procedures are and to address them as early as possible, preferably as part of the initial proposal development. When seeking cooperation, it is always helpful to

offer to provide information to the participants on what is learned, either through personal feedback or a workshop in which findings can be discussed. If this is too time consuming, a copy of the report or executive summary may well do. The main idea here is to provide incentives for people or organizations to take the time to participate in your evaluation.

Second, the needs of the participants must be considered. Being part of an evaluation can be very threatening to participants, and they should be told clearly and honestly why the data are being collected and how the results will be used. On most survey-type studies, assurances are provided that no personal repercussions will result from information presented to the evaluator and, if at all possible, individuals and their responses will not be publicly associated in any report. This guarantee of anonymity frequently makes the difference between a cooperative and a recalcitrant respondent.

*Participants should be told clearly and honestly why the data are being collected and how the results will be used.*

There may, however, be some cases when identification of the respondent is deemed necessary, perhaps to enforce the credibility of an assertion. In studies that use qualitative methods, it may be more difficult to report all findings in ways that make it impossible to identify a participant. In qualitative studies, the number of respondents is often quite small, especially if one is looking at respondents with characteristics that are of special interest in the analysis (for example, older teachers or teachers who hold graduate degrees). Thus, even if a finding does not name the respondent, it may be possible for someone (a colleague, an administrator) to identify a respondent who made a critical or disparaging comment in an interview. In such cases, the evaluation should include a step wherein consent is obtained before including such information. Consent may also be advisable where a sensitive comment is reported, despite the fact that the report itself includes no names. Common sense is the key here. The American Evaluation Association (AEA) has a set of guiding principles for evaluators (AEA, 2005) that provide some very important tips in this area under the heading "Respect for People."

Third, data collectors must be carefully trained and supervised, especially where multiple data collectors are used. This training should include providing the data collectors with information about the culture and rules of the community in which they will be interacting (especially if the community differs from that of the data collector) as well as technical information. It is important that data collectors understand the idiom of those with whom they will be interacting so that two-way communication and understanding can be maximized.

*Periodic checks need to be carried out to make sure that well-trained data collectors do not "drift" away from the prescribed procedures over time.*

The data collectors must be trained so that they all see things in the same way, ask the same questions, and use the same prompts. It is important to establish inter-rater reliability: when ratings or categorizations of data collectors for the same event are compared, an inter-rater reliability of 80 percent or more is

desired. Periodic checks need to be conducted to make sure that well-trained data collectors do not "drift" away from the prescribed procedures over time. Training sessions should include performing the actual task (extracting information from a database, conducting an interview, performing an observation), role-playing (for interviews), and comparing observation records of the same event by different observers.

When the project enters a new phase (for example, when a second round of data collection starts), it is usually advisable to schedule another training session and to check inter-rater reliability again. If funds and technical resources are available, other techniques (for example, videotaping of personal interviews or recording of telephone interviews) can also be used for training and quality control after permission has been obtained from participants.

Evaluations need to include procedures to guard against possible distortion of data because of well intended but inappropriate "coaching" of respondents—an error frequently made by inexperienced or overly enthusiastic staff. Data collectors must be warned against providing value-laden feedback to respondents or engaging in discussions that might well bias the results. One difficult but important task is understanding one's own biases and making sure that they do not interfere with the work at hand. This is a problem all too often encountered when dealing with volunteer data collectors, such as parents in a school or teachers in a center, or teaching assistants at a university. They volunteer because they are interested in the project that is being evaluated or are advocates for or critics of it. Unfortunately, the data they produce may reflect their own perceptions of the project, as much as or more than that of the respondents, unless careful training is undertaken to avoid this "pollution." Bias or perceived bias may compromise the credibility of the findings and the ultimate use to which they are put. An excellent source of information on these issues is the section on accuracy standards in *The Program Evaluation Standards* (Joint Committee on the Standards for Educational Evaluation, 2010).

Fourth, try to get data from as many members of your sample as possible. The validity of your findings depends not only on how you select your sample, but also on the extent to which you are successful in obtaining data from those you have selected for study. It is important to follow up with individuals who are nonresponsive to the initial contact to try to get them to participate to avoid bias in your respondent sample. This can mean sending surveys out two to three times or rescheduling interviews or observations on multiple occasions. Newcomer and Triplett (2004) say that nonresponse is usually a concern. They say that a response rate of 70 percent or higher is considered to be high quality and recommend some adjustment, such as weighting, for response rates between 50 and 70 percent. Wherever possible, assessing whether there is some systematic difference between those who respond and those who do not is always advisable. If differences are found, they

*It is important to follow up with individuals who are nonresponsive to the initial contact to try to get them to participate.*

should be noted and the impact on the generalizability of findings recorded.

An important, but often ignored, step is gathering data from those who may have been part of a treatment initially but dropped out along the way. Following up on these ex-participants provides a fuller picture of impacts on the treatment group, as well as an assessment of the impacts of dropping out.

Finally, the data gathering should cause as little disruption as possible. Among other things, this means being sensitive to the schedules of the people or the project. It also may mean changing approaches as situations come up. For example, instead of asking a respondent to provide data on the characteristics of project participants—a task that may require considerable time for the respondent to pull the data together and develop summary statistics—the data collector may need to work from raw data, applications, monthly reports, etc., and personally do the compilation.

## Analyzing the Data

Once the data are collected, they must be analyzed and interpreted. The steps followed in preparing the data for analysis and interpretation differ, depending on the type of data. The interpretation of qualitative data may in some cases be limited to descriptive narratives, but other qualitative data may lend themselves to systematic analyses through the use of quantitative approaches such as thematic coding or content analysis. Analysis includes several steps:

- Check the raw data and prepare them for analysis.
- Conduct initial analysis based on the evaluation plan.
- Conduct additional analyses based on the initial results.
- Integrate and synthesize findings.

The first step in quantitative data analysis is checking data for responses that may be out of line or unlikely. Such instances include selecting more than one answer when only one can be selected, always choosing the third alternative on a multiple-choice test of science concepts, reporting allocations of time that add up to more than 100 percent, giving inconsistent answers, etc. Where such problematic responses are found, it may be necessary to eliminate the item or items from the data to be analyzed.

After this is done, the data are prepared for computer analysis; usually this involves coding and entering (keying or scanning) the data with verification and quality control procedures in place.

The next step is to carry out the data analysis specified in the evaluation plan. While new information gained as the evaluation evolves may well cause some analyses to be added or subtracted, it is a good idea to start with the set of analyses that seemed originally to be of interest. Statistical programs are available on easily accessible software that make the data analysis task considerably easier today than it was 25 years ago. Analysts still need to be careful, however, that the data sets they are using meet the assumptions of the technique being used. For example, in the analysis of quantitative data, different approaches may be used to analyze continuous data as opposed to categorical data. Using an incorrect technique can result in invalidation of the whole evaluation project. Increasingly, computerized systems for qualitative analysis are being used to manage the large sets of narrative data. These provide support to the analyst and a way of managing the large amounts of data that are typically collected (but do not eliminate the need for careful analysis and decision making on the part of the evaluator).

It is very likely that the initial analyses will raise as many questions as they answer. The next step, therefore, is conducting a second set of analyses to further address these questions. If, for example, the first analysis looked at overall teacher performance, a second analysis might subdivide the total group into subunits of particular interest—e.g., more experienced versus less experienced teachers; teachers rated very successful by mentors versus teachers rated less successful—and examine whether any significant differences were found between them. These reanalysis cycles can go through several iterations as emerging patterns of data suggest other interesting avenues to explore. Sometimes the most intriguing results emerge from the data; they are ones that were not anticipated or sought despite the thoroughness of your initial analysis plan. In the end, it becomes a matter of balancing the time and money available against the inquisitive spirit in deciding when the analysis task is completed.

*It is very likely that the initial analyses will raise as many questions as they answer.*

It should be noted that we have not attempted to go into any detail on the different statistical techniques that might be used for quantitative analysis. Indeed, this discussion is the subject of many books and textbooks. Suffice it to say that most evaluations rely on fairly simple descriptive statistics—means, frequencies, etc. However, where more complex analyses and causal modeling are derived, evaluators will need to use analyses of variance, regression analysis, Hierarchical Linear Modeling, or for structural equation modeling.

The final task is to present the results of the varied analyses, to integrate the separate analyses into an overall picture, and to develop conclusions regarding what the data show. Sometimes this integration of findings becomes very challenging, as the different data sources do not yield

completely consistent findings. While it is preferable to be able to produce a report that reconciles differences and explains the apparent contradictions, sometimes the findings must simply be allowed to stand as they are, unresolved and, it is hoped, thought provoking.

## Reporting the Findings

The next stage of the project evaluation is reporting what has been found. This requires pulling together the data collected, distilling the findings in light of the questions the evaluation was originally designed to address, and disseminating the findings.

Formal reports developed by evaluators typically include six major sections:

- Background
- Evaluation study questions
- Evaluation procedures
- Data analyses
- Findings
- Conclusions (and recommendations)

### Background

The background section describes (1) the problem or needs addressed, (2) a literature review, if relevant, (3) the stakeholders and their information needs, (4) the participants, (5) the project's objectives, (6) the activities and components, (7) location and planned longevity of the project, (8) the resources used to implement the project, and (9) the project's expected measurable outcomes.

Notable constraints that existed in what the evaluation was able to do are also pointed out in this section. For example, it may be important to point out that conclusions are limited by the fact that no appropriate comparison group was available or that only the short-term effects of program participation could be examined.

### Evaluation Study Questions

An evaluation is based on the need for specific information, and stakeholders, such as Congress, NSF-funded program and project directors, and the participants, have somewhat different information

needs. There are many questions to be asked about a project, and they cannot be answered at one time. This section of the report describes the questions that the study addressed. As relevant, it also points out some important questions that could not be addressed because of factors such as time, resources, or inadequacy of available data collection techniques.

**Evaluation Procedures**

This section of the report describes the groups that participated in the evaluation study. It describes who these groups were and how the particular sample of respondents included in the study was selected from the total population available, if sampling was used. Important points noted are how representative the sample was of the total population, whether the sample volunteered (self-selected) or was chosen using some sampling strategy by the evaluator, and whether or not any comparison or control groups were included. If comparison groups were included, it is important to provide data attesting to their equivalence or indicate how the problem of imperfect equivalence was addressed.

This section also describes the types of data collected and the instruments used for the data collection activities. For example, they could be:

- Data for identified critical indicators, e.g., grades for specific subjects, grade point averages (GPAs)
- Ratings obtained in questionnaires and interviews designed for project directors, students, faculty, and graduate students
- Descriptions of classroom activities from observations of key instructional components of the project
- Examinations of extant data records, e.g., letters, planning papers, and budgets

It is helpful at the end of this section to include a matrix or table that summarizes the evaluation questions, the variables, the data collection approaches, the respondents, and the data collection schedule.

**Data Analyses**

This section describes the techniques used to analyze the data that were collected. It describes the various stages of analysis that were implemented and the checks that were carried out to make sure that the data were free of as many confounding factors as possible. Frequently, this section contains a discussion of the techniques used to make sure that the sample of participants that actually participated in the study was, in fact, representative of the population from which it came. Any

limitations in the generalizability of findings are noted. (That is, there is sometimes an important distinction between the characteristics of the sample that was selected for participation in the evaluation study and the characteristics of those who actually participated, were retained, returned questionnaires, attended focus groups, etc.)

Again, a summary matrix is a very useful illustrative tool.

**Findings**

This section presents the results of the analyses described previously. The findings are usually organized in terms of the questions presented in the section on evaluation study questions. Each question is addressed, regardless of whether or not a satisfactory answer can be provided. It is just as important to point out where the data are inconclusive as where the data provide a positive or negative answer to an evaluation question. Visuals such as tables and graphical displays are an appropriate complement to the narrative discussion. As findings are presented, it is important to make clear any limitations in the work that affect its validity. For example, if there is unequal attrition in the treatment and control groups, it should be noted and the implications stated.

At the end of the findings section, it is helpful to have a summary that presents the major conclusions. Here, "major" is defined in terms of both the priority of the question in the evaluation and the strength of the finding from the study. However, the summary of findings would always include a statement of what was learned with regard to outcomes, regardless of whether the data were conclusive.

**Conclusions (and Recommendations)**

The conclusions section reports the findings with more broad-based and summative statements. These statements must relate to the findings of the project's evaluation questions and to the goals of the overall program. Sometimes the conclusions section goes a step further and includes recommendations either for NSF or for others undertaking projects similar in goals, focus, and scope. Care must be taken to base any recommendations solely on robust findings that are data based, and not on anecdotal evidence, no matter how appealing.

**Other Sections**

In addition to these six major sections, formal reports also include one or more summary sections. These might be:

- An abstract: a summary of the study and its findings presented in approximately one-half page of text.

- An executive summary: a summary, which may be as long as four to 10 pages, that provides an overview of the evaluation, its findings, and implications. Sometimes the executive summary also serves as a nontechnical digest of the evaluation report.

**How Do You Develop an Evaluation Report?**

Although we usually think about report writing as the last step in an evaluation study, a good deal of the work actually can and does take place by your evaluators before the project is completed. The background section, for example, can be based largely on the original evaluation design document. While there may be some events that cause minor differences between the study as planned and the study as implemented, the large majority of information, such as research background, the problem addressed, the stakeholders, and the project's goals, will remain essentially the same. Reports that are simply written technical documents are no longer acceptable; successful reporting involves giving careful thought to the creation and presentation of the information in ways that will be accessible to broad lay audiences, as well as to professional audiences. Derivative, nontechnical summaries, as well as electronic media, are becoming increasingly important means of sharing information.

For example, many agencies share information broadly by putting it on the web. Sometimes information is posted on a CD-ROM, which allows large amounts of information—including copies of instruments, data sets, and other technical analyses—as well as the written report to be contained on a small, easy-to-access carrier. In addition, electronic tools can be used to make colorful, clear, attention-getting presentations about a study and its findings.

If there is a written evaluation design, the material in this design can be used for the section on evaluation study questions and sample, data collection, and instrumentation. The data analysis section is frequently an updated version of what was initially proposed. However, as we noted earlier, data analysis can take on a life of its own, as new ideas emerge when data are explored. The final data analysis may be far different than what was initially envisioned.

The findings and conclusions sections are the major new sections to be written at the end of an evaluation study. These may present somewhat of a challenge because of the need to balance comprehensiveness with clarity, and rigorous, deductive thinking with intuitive leaps. One of the errors frequently made in developing a findings section is what we might call the attitude of "I analyzed it, so I am going to report it." That is, evaluators may feel compelled to report analyses that at first appeared fruitful but ultimately resulted in little information of interest. In most cases, it is sufficient to note that these analyses were conducted and that the results were inconclusive. Presentation of tables showing that no differences occurred or no patterns emerged is probably not a good idea

unless there is a strong conceptual or political reason for doing so. Even in the latter case, it is prudent to note the lack of findings in the text and to provide the backup evidence in appendices or some technical supplement.

One tip to follow when writing these last sections is to ask colleagues or stakeholders to review what you have written and provide feedback before the report reaches its final form. These reviewers can assist in assessing the clarity and completeness of what you have written, as well as providing another set of eyes to examine your arguments and, possibly, challenge your interpretations. It is sometimes very hard to get enough distance from your own analyses after you have been immersed in them.

Finally, the information needs to be provided in a manner and style that is appropriate, appealing, and compelling to the person being informed. For example, a detailed numerical table with statistical test results might not be the best way to provide a school board member with achievement data on students. On the other hand, technical audiences want detailed information about what was done and what was found. Different reports may have to be provided for the different audiences, and it may well be that a written report is not even the preferred alternative. Written reports are frequently accompanied by other methods of communicating findings, such as PowerPoint presentations or web-based documents in full or shortened form. Still, the formal, technical report remains the primary way of communicating evaluation findings, and a sample outline for such a document is presented in Exhibit 12.

It should be noted that while discussions of communicating study results generally stop at the point of presenting a final report of findings, there are important additional steps that should be considered. Especially when a new product or practice turns out to be successful, as determined by a careful evaluation, dissemination is an important next step. Planning for dissemination is important and can be as challenging as the evaluation itself.

## Disseminating the Information

The final stage in project evaluation is dissemination. Ideally, planning for dissemination begins in the early stages of developing a project, with audiences and their needs for information determined simultaneously with project design. It is useful to make a listing of the various audiences with whom you would like to share findings. The listing may be very similar to those included in your stakeholder group and would include:

- The funding source(s)
- Potential funding sources

**Exhibit 12.—Formal report outline**

I. Summary sections
   - A. Abstract
   - B. Executive summary

II. Background
   - A. Problems or needs addressed
   - B. Literature review
   - C. Stakeholders and their information needs
   - D. Participants
   - E. Project's objectives
   - F. Activities and components
   - G. Location and planned longevity of the project
   - H. Resources used to implement the project
   - I. Project's expected measurable outcomes
   - J. Constraints

III. Evaluation study questions
   - A. Questions addressed by the study
   - B. Questions that could not be addressed by the study (when relevant)

IV. Evaluation procedures
   - A. Sample
     1. Selection procedures
     2. Representativeness of the sample
     3. Use of comparison or control groups, if applicable
   - B. Data collection
     1. Methods
     2. Instruments
   - C. Summary matrix
     1. Evaluation questions
     2. Variables
     3. Data gathering approaches
     4. Respondents
     5. Data collection schedule

V. Findings
   - A. Results of the analyses organized by study question

VI. Conclusions
   - A. Broad-based, summative statements
   - B. Recommendations, when applicable

- Others involved with similar projects or areas of research

- Community members, especially those who are directly involved with the project or might be involved

- Members of the business or political community, etc.

In developing a dissemination approach, two areas need to be considered: what these various groups need to know, and the best manner for communicating information to them. For example, NSF will want both a formal final report with technical details and an executive summary with highlights of the findings. This report should link your project to NSF's overall goals for the program and show how what you accomplished informs or relates to these goals. It is also important to identify contributions to the overall research or knowledge base in your area of investigation. Keep in mind NSF's strategic outcomes discussed in Chapter 1, as identified in GPRA, as you develop your report.

A report to the community that is directly involved, or might be involved, would be presented in a less formal and detailed fashion, with a minimum of technical detail. This report could take many forms, e.g., a newsletter, a fact sheet, or even a short journalistic article. In-person presentations in which interactive discussion can occur may be especially useful. In developing a report for this group, it is important both to share the results and to help these stakeholders understand what the results mean for them and what they might do with the information.

Newcomer and Wirtz (2004) provide some good tips on reporting to officials. They caution against using language that is too technical and hedging on what the data mean by presenting too many caveats about the possibly tentative nature of the statistical results. Further, they advise, "A distinction between statistical and practical importance may be too much to provide to high level decision makers. Instead, only findings that are of practical importance should be presented" (p. 461).

If your work is successful and you have a product to share, such as a module for instruction, other strategies may be used. At a minimum, presentations at conferences and meetings will increase awareness of your work and may cause others to build on or adopt your product. More formally, it may be useful to seek support to package your product for others to use along with support materials and even training workshops.

Although the idea of dissemination is most frequently associated with instances where projects have "worked" (with what this means differing depending on the context of the project), it is also important to share results in instances where hypotheses have not been supported or well-constructed attempts at innovation have not proven fruitful. Such knowledge is probably most relevant to your funders and your colleagues in the research world and can be shared through professional communications.

**References**

American Evaluation Association. (2010). *Guiding Principles for Evaluators.* Guiding Principles brochure.

Joint Committee on the Standards for Educational Evaluation. (2010). *The Program Evaluation Standards*. 2nd Ed. Thousand Oaks, CA: Sage Publications.

Newcomer, K., and Triplett, T. (2004). Using Surveys. In *Handbook of Practical Program Evaluation: Second Edition*, edited by J. Wholey, H. Hatry, and K. Newcomer. San Francisco, CA: Jossey-Bass.

Newcomer, K., and Wirtz, P. (2004). Using Statistics in Evaluation. In *Handbook of Practical Program Evaluation: Second Edition*, edited by J. Wholey, H. Hatry, and K. Newcomer. San Francisco, CA: Jossey-Bass.

Chapter 5

# DATA COLLECTION METHODS: SOME TIPS AND COMPARISONS

In the previous chapter, we identified two broad types of evaluation methodologies: quantitative and qualitative. In this section, we talk more about the debate over the relative virtues of these approaches and discuss some of the advantages and disadvantages of different types of instruments. In such a debate, two types of issues are considered: theoretical and practical.

## Theoretical Issues

Most often these center on one of three topics:

- The value of the types of data
- The relative scientific rigor of the data
- Basic, underlying philosophies of evaluation

### Value of the Data

Quantitative and qualitative techniques provide a tradeoff between breadth and depth, and between generalizability and targeting to specific (sometimes very limited) populations. For example, a quantitative data collection methodology such as a sample survey of high school students who participated in a special enrichment in nanotechnology program can yield representative and broadly generalizable information about the proportion of participants who plan to major in science when they get to college and how this proportion differs by gender. But at best, the survey can elicit only a few, often superficial reasons for this gender difference. On the other hand, separate focus groups (a qualitative technique related to a group interview) conducted with small groups of men and women students will provide many more clues about gender differences in the choice of science majors, and the extent to which the nanotechnology program changed or reinforced attitudes. The focus group technique is, however, limited in the extent to which findings apply beyond the specific individuals included in the groups.

### Scientific Rigor

Data collected through quantitative methods are often believed to yield more objective and accurate information because they were collected using standardized methods, can be replicated, and, unlike qualitative

data, can be analyzed using sophisticated statistical techniques. In line with these arguments, traditional wisdom has held that qualitative methods are most suitable for formative evaluations, whereas summative evaluations require "hard" (quantitative) measures to judge the ultimate value of the project.

This distinction is too simplistic. Both approaches may or may not satisfy the canons of scientific rigor. Quantitative researchers are becoming increasingly aware that some of their data may not be accurate and valid, because the respondents may not understand the meaning of questions to which they respond, either because people's recall of events is often faulty, or because critical control variables were not included in the analyses. On the other hand, qualitative researchers have developed better techniques for classifying and analyzing large bodies of descriptive data. It is also increasingly recognized that all data collection—quantitative and qualitative—operates within a cultural context and is affected to some extent by the perceptions and beliefs of investigators and data collectors.

**Philosophical Distinction**

*Researchers and scholars differ about the respective merits of the two approaches, largely because of different views about the nature of knowledge and how knowledge is best acquired.*

Researchers and scholars differ in their opinions about the respective merits of the two approaches, largely because of different views about the nature of knowledge and how knowledge is best acquired. Clark and Creswell (2008) qualitative researchers feel that there is no objective social reality and all knowledge is "constructed" by observers who are the product of traditions, beliefs, and the social and political environments within which they operate. Quantitative researchers, who also have abandoned naive beliefs about striving for absolute and objective truth in research, continue to adhere to the scientific model and to develop increasingly sophisticated statistical techniques to measure social phenomena.

This distinction affects the nature of research designs. According to its most orthodox practitioners, qualitative research does not start with clearly specified research questions or hypotheses to be tested; instead, questions are formulated after open-ended field research has been completed (Lofland and Lofland, 1995) This approach is difficult for program and project evaluators to adopt, since specific questions about the effectiveness of interventions being evaluated are expected to guide the evaluation. Some researchers have suggested that a distinction be made between Qualitative work and qualitative work: Qualitative work (large Q) involves participant observation and ethnographic field work, whereas qualitative work (small q) refers to open-ended data collection methods such as in-depth interviews embedded in structured research (Kidder and Fine, 1987). The latter are more likely to meet NSF evaluation needs.

## Practical Issues

On the practical level, four issues can affect the choice of method:

- Credibility of findings
- Staff skills
- Costs
- Time constraints

### Credibility of Findings

Evaluations are designed for various audiences, including funding agencies, policymakers in governmental and private agencies, project staff and clients, researchers in academic and applied settings, and various other stakeholders. Experienced evaluators know that they often deal with skeptical audiences or stakeholders who seek to discredit findings that are too critical or not at all critical of a project's outcomes. For this reason, an evaluation methodology may be rejected as unsound or weak for a specific case.

The major stakeholders for NSF projects are policymakers within NSF and the federal government, state and local officials, and decision makers in the educational community where the project is located. In most cases, decision makers at the national level favor quantitative information because these policymakers are accustomed to basing funding decisions on numbers and statistical indicators. On the other hand, many stakeholders in the educational community are often skeptical about statistics and "number crunching" and consider the richer data obtained through qualitative research to be more trustworthy and informative. A particular case in point is the use of traditional test results, a favorite outcome criterion for policymakers, school boards, and parents, but one that teachers and school administrators tend to discount as a minimalistic tool for assessing true student learning.

### Staff Skills

Qualitative methods, including in-depth interviewing, observations, and the use of focus groups, require good staff skills and considerable training and monitoring to yield trustworthy data. Some quantitative research methods can be mastered easily with the help of simple training manuals; this is true of small-scale, self-administered questionnaires in which most questions can be answered by yes/no checkmarks or selecting numbers on a simple scale. Large-scale, complex studies, however, usually require more skilled personnel to design the study,

develop the instruments, manage data collection, and ensure the integrity of the analysis.

**Costs**

It is difficult to generalize about the relative costs of the two methods; much depends on the amount of information needed, quality standards followed for the data collection, and the number of cases required for reliability and validity. A true experiment, with participants randomized into treatment and control groups, will be expensive, especially if the participants are followed over time. A small study, using a short survey consisting of a few "easy" questions, would be inexpensive, but it also would provide only limited data. Even cheaper would be substituting a focus group session for a subset of 25–50 participants. While this latter method might be less costly, the data would be primarily useful for generating new hypotheses to be tested by more appropriate quantitative or qualitative methods. To obtain robust findings, the cost of data collection is bound to be high regardless of method.

**Time Constraints**

Similarly, data complexity and quality affect the time needed for data collection and analysis. Although technological innovations have shortened the time needed to process quantitative data, a good evaluation requires considerable time to design and implement. When true experiments are used, additional time must be set aside to recruit and screen subjects for the treatment and control groups. Resources must also be set aside to keep participants on board during the implementation of the study to reduce attrition. Tracking is needed when participants drop out. However, qualitative methods may be even more time consuming because data collection and data analysis overlap, and the process encourages the exploration of new evaluation questions. If insufficient time is allowed for evaluation, it may be necessary to curtail the amount of data to be collected or to cut short the analytic process, thereby limiting the value of the findings. For evaluations that operate under severe time constraints—for example, where budgetary decisions depend on the findings—choosing the best method can present a serious dilemma.

*For evaluations that operate under severe time constraints—for example, where budgetary decisions depend on the findings—choosing the best method can present a serious dilemma.*

The debate with respect to the merits of qualitative versus quantitative methods is still ongoing in the academic community, but when it comes to the choice of methods in conducting project evaluations, a pragmatic strategy has been gaining increased support. Respected practitioners have argued for integrating the two approaches by putting together packages of the available imperfect methods and theories, which will minimize biases by selecting the least biased and most appropriate method for each evaluation subtask (Shadish, 1993). Others have stressed the advantages

of linking qualitative and quantitative methods when performing studies and evaluations, showing how the validity and usefulness of findings will benefit from this linkage (Miles and Huberman, 1994).

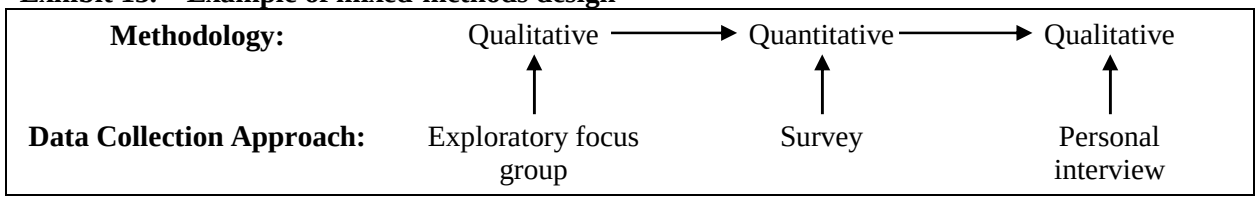
## Using the Mixed-Methods Approach

While quantitative data are always needed if a case is to be made for the efficacy of an intervention or approach, we feel that a strong case can be made for including qualitative elements in the great majority of evaluations of NSF projects. To ignore the complexity of the background is to impoverish the evaluation. Similarly, when investigating human behavior and attitudes, it is most fruitful to use a variety of data collection methods. By using different sources and methods at various points in the evaluation process, the evaluation team can build on the strength of each type of data collection and minimize the weaknesses of any single approach (Kane and Trochim, 2007). A mixed-methods approach to evaluation can increase both the validity and the reliability of evaluation data.

*A strong case can be made for including qualitative elements in the great majority of evaluations of NSF projects.*

The range of possible benefits that carefully designed mixed-methods designs can yield has been conceptualized by a number of evaluators. The validity of results can be strengthened by using more than one method to study the same phenomenon. This approach—called triangulation—is most often mentioned as the main advantage of the mixed-methods approach. Combining the two methods pays off in improved instrumentation for all data collection approaches and in sharpening the evaluator's understanding of findings. A typical design might start out with a qualitative segment such as a focus group discussion alerting the evaluator to issues that should be explored in a survey of program participants, followed by the survey, which in turn is followed by indepth interviews to clarify some of the survey findings (Exhibit 13).

**Exhibit 13.—Example of mixed-methods design**




It should be noted that triangulation, while very powerful when sources agree, can also pose problems for the analyst when different sources yield different, even contradictory information. There is no formula for resolving such conflicts, and the best advice is to consider disagreements in the context in which they emerge. Some suggestions for resolving differences are provided by Altshuld and Witkin (2000).

This sequential approach is only one of several that evaluators might find useful. Thus, if an evaluator has identified subgroups of program participants or specific topics for which in-depth information is needed, a limited qualitative data collection can be initiated while a more broad-based survey is in progress.

Mixed methods may also lead evaluators to modify or expand the adoption of data collection methods. This can occur when the use of mixed methods uncovers inconsistencies and discrepancies that should alert the evaluator to the need for re-examining data collection and analysis procedures. The philosophy guiding the suggestions outlined in this Handbook can be summarized as follows:

> The evaluator should attempt to obtain the most useful information to answer the critical questions about the project and, in so doing, rely on a mixed-methods approach whenever possible.

This approach reflects the growing consensus among evaluation experts that both qualitative and quantitative methods have a place in the performance of effective evaluations, be they formative or summative.

**References**


Altshuld, J., and Witkin, B.R. (2000). *Transferring Needs into Solution Strategies*. Newbury Park, CA: Sage.

Clark, V.L., Plano, and Creswell, J.W. (eds). (2008). *The Mixed Method Reader.* Thousand Oaks, CA: Sage Publications.

Kane, M., and Trochim, W. (2007). *Concept Mapping for Planning and Evaluation.* Thousand Oaks, CA: Sage Publications.

Kidder, L., and Fine, M. (1987). *Qualitative and Quantitative Methods: When Stories Converge. Multiple Methods in Program Evaluation*. New Directions for Program Evaluation, No. 35. San Francisco, CA: Jossey-Bass.

Lofland, J., and Lofland, L.H. (1995). *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont, CA: Wadsworth Publishing Company.

Miles, M.B., and Huberman, A.M. (1994). *Qualitative Data Analysis.* 2nd Ed. Newbury Park, CA: Sage.

Shadish, W.R. (1993) *Program Evaluation: A Pluralistic Enterprise*. New Directions for Program Evaluation, No. 60. San Francisco, CA: Jossey-Bass.

# Chapter 6 REVIEW AND COMPARISON OF SELECTED TECHNIQUES

In this section, we describe and compare the most common quantitative and qualitative methods employed in project evaluations. These include surveys, in-depth interviews, focus groups, observations, and tests. We also review briefly some other less frequently used qualitative techniques. Advantages and disadvantages are summarized. For those interested in learning more about data collection methods, a list of recommended readings is provided at the end of the report. Readers may also want to consult the Online Evaluation Resource Library (OERL) website (http://oerl.sri.com), which provides information on approaches used in NSF project evaluations, as well as reports, modules on constructing designs, survey questionnaires, and other instruments.

## Surveys

Surveys are a very popular form of data collection, especially when gathering information from large groups, where standardization is important. Surveys can be constructed in many ways, but they always consist of two components: questions and responses. While sometimes evaluators choose to keep responses "open ended" (i.e., allow respondents to answer in a free-flowing narrative form), most often the "close-ended" approach in which respondents are asked to select from a range of predetermined answers is adopted. Open-ended responses may be difficult to code and require more time and resources to handle than close-ended choices. Responses may take the form of a rating on some scale (e.g., rate a given statement from one to four on a scale from "agree" to "disagree"), may give categories from which to choose (e.g., select from potential categories of partner institutions with which a program could be involved), or may require estimates of numbers or percentages of time in which participants might engage in an activity (e.g., the percentage of time spent on teacher-led instruction or cooperative learning).

Although surveys are popularly referred to as paper-and-pencil instruments, this too is changing. Evaluators are increasingly using methods that take advantage of the emerging technologies. Thus, surveys may be administered via computer-assisted calling, as e-mail attachments, and as web-based online data collection systems.

Selecting the best method for collecting surveys requires weighing a number of factors. These included the complexity of questions, resources available, the project schedule, the intended audience, etc. For example, web-based surveys are attractive for a number of reasons. First, because the data collected can be put directly into a database, the time and steps between data collection and analysis can be shortened. Second, it is possible to build in checks that keep out-of-range responses from being

entered. For some populations, however, access to computers may still be more limited. Many projects using surveys thus combine web-based and traditional paper-and-pencil approaches.

### When to Use Surveys

Surveys are typically selected when information is to be collected from a large number of people or when answers are needed to a clearly defined set of questions. Surveys are good tools for obtaining information on a wide range of topics when in-depth probing of responses is not necessary, and they are useful for both formative and summative purposes. Frequently, the same survey is used at spaced intervals of time to measure progress along some dimension or change in behavior. Exhibit 14 shows the advantages and disadvantages of surveys.

**Exhibit 14.—Advantages and disadvantages of surveys**

**Advantages:**

- Good for gathering descriptive data
- Can cover a wide range of topics
- Are relatively inexpensive to use
- Can be analyzed using a variety of existing software

**Disadvantages:**

- Self-report may lead to biased reporting
- Data may provide a general picture but lack depth
- May not provide adequate information on context

## Interviews

The use of interviews as a data collection method begins with the assumption that the participants' perspectives are meaningful, knowable, and can be made explicit, and that their perspectives affect the success of the project. An in-person or telephone interview, rather than a paper-and-pencil survey, is selected when interpersonal contact is important and when opportunities for follow-up of interesting comments are desired.

Two types of interviews are used in evaluation research: structured interviews, in which a carefully worded questionnaire is administered, and in-depth interviews, in which the interviewer does not follow a rigid form. In the former, the emphasis is on obtaining answers to carefully phrased questions. Interviewers are trained to deviate only minimally from the question wording to ensure uniformity of interview administration. In the latter, however, the interviewers seek to encourage free and open responses, and there may be a tradeoff between comprehensive coverage of topics and in-depth exploration of a more limited set of questions. In-depth interviews also encourage capturing respondents' perceptions in their own words, a very desirable strategy in qualitative data collection. This technique allows the evaluator to present the meaningfulness of the experience from the respondent's perspective. In-depth interviews are conducted with individuals or small groups of individuals.

**When to Use Interviews**

Interviews can be used at any stage of the evaluation process. In-depth interviews are especially useful in answering questions such as those suggested by Patton (1990):

- What does the program look and feel like to the participants? To other stakeholders?
- What do stakeholders know about the project?
- What thoughts do stakeholders knowledgeable about the program have concerning program operations, processes, and outcomes?
- What are participants' and stakeholders' expectations?
- What features of the project are most salient to the participants?
- What changes do participants perceive in themselves as a result of their involvement in the project?

Specific circumstances for which in-depth interviews are particularly appropriate include situations involving complex subject matter, detailed information, high-status respondents, and highly sensitive subject matter. Exhibit 15 shows the advantages and disadvantages of interviews.

**Exhibit 15.—Advantages and disadvantages of interviews**

**Advantages:**

- Usually yield richest data, details, new insights
- Permit face-to-face contact with respondents
- Provide opportunity to explore topics in depth
- Allow interviewer to experience the affective as well as cognitive aspects of responses
- Allow interviewer to explain or help clarify questions, increasing the likelihood of useful responses
- Allow interviewer to be flexible in administering interview to particular individuals or in particular circumstances

**Disadvantages**:

- Expensive and time consuming
- Need well-qualified, highly trained interviewers
- Interviewee may distort information through recall error, selective perceptions, desire to please interviewer
- Flexibility can result in inconsistencies across interviews
- Volume of information very large; may be difficult to transcribe and reduce data

## Social Network Analysis

Like case studies, social network data may be derived from a variety of methods and sources. However, while the underlying methods of data collection are common, the models and methods of social network analysis (SNA) present researchers with a number of unique challenges and opportunities. Specifically, SNA-based methods focus on the quantification, description, and analysis of the patterns or structure of dyadic (e.g., pairwise) social ties and interdependencies that occur in social groups and organizations. Thus, SNA is often used to address program evaluation questions that involve questions about relational ties, as either an independent or dependent variable (or both), and their relationship to the characteristics (e.g., individual differences, demographics, training, etc.) of individuals and groups (termed *actors*) within the network (Durland and Fredericks, 2005).

Examples of social networks might include knowledge-sharing relationships among teachers, friendship relations among students, joint

attendance at training sessions, or even patterns of student flows between regions or institutions. Common research questions might include whether individuals with certain characteristics are more or less likely to share certain relationships; whether individuals who share relationships are more or less likely to have similar characteristics; whether people who share one type of relationships are more or less likely to share other relationships; whether relationships in the group are sparse or dense, centralized or decentralized, and cohesive or fragmented; and how these relationships change and develop over time.

When studying social networks, researchers first need to decide how to define or identify the boundaries of the network (e.g., who the actors in your network are), how they will collect the data, and what kind of network data they wish to study. Most models and methods for network analysis are geared toward the analysis of *sociocentric* data, where researchers try to map out the complete pattern of relationships between all members of a given group. This type of data is the most difficult to collect, but provides the greatest amount of information on a group's social structure, and the most flexibility in terms of analysis. Researchers may also study *egocentric* networks; in egocentric data, instead of collecting data on the entire network, network data are collected only from a sample of individuals, each of whom provides information about their personal network. Egocentric data are much easier to collect, but also much more limited in the kinds of research questions that can be addressed. However, other types of network data exist; these include *two-mode* networks (which involve relationships between two different types of actors, like which individuals attended the same social events) and cognitive social structures (which involve individual perceptions of sociocentric networks).

Data may be collected using a variety of different methods, such as direct observation, interviews, questionnaires, or archival data (such as e-mail records). Questionnaires tend to be the most widely used method; these require the researcher to make choices about how to elicit responses about the social network—for example, whether to use free recall (where respondents are simply asked to write down names) or roster-based methods (where respondents are provided a full list of names and asked to select those with whom they have ties). Researchers should be aware that the effects of missing data in network studies can be especially problematic for many types of research questions, and that ways to handle missing data in sociocentric networks are not yet well developed. Thus, every effort should be made to minimize missing data; some researchers try to ensure response rates of 80 percent or greater.

Measures, models, and methods specifically designed for social network data are typically required in order to address SNA-related research questions. Fortunately, there are a number of powerful tools available for analysis and visualization; many of these are free or low cost, widely used, and well documented. These include UCINET, the most widely used software for SNA; Pajek, which is able to handle very large

networks; and the SNA package for the open-source statistics program R, which is very flexible but has a high learning curve.

### When to Use Social Network Analysis

Social network analysis is especially appropriate when research questions involve quantifying or describing patterns of dyadic (e.g., pairwise) social ties within a group or population, or linking these ties to the characteristics and outcomes of individuals and groups. For example, SNA might be indispensable for a researcher studying the extent to which a group's communication network is cohesive or fragmented, whether certain types of individuals are more likely to be central or peripheral to the communications network, and whether an individual's position in the communication structure is related to attitudes and behaviors like job satisfaction and turnover intentions. In addition to testing more formal research questions, it can also be used for diagnostic purposes, such as identifying social groups in a school that do not communicate with one another.

### Advantages and Disadvantages of SNA

Advantages

- Allows for identifying and quantifying structural patterns in interpersonal and interorganizational relationships
- Can be used for exploratory and diagnostic purposes
- Many important concepts in SNA (such as centrality) are easily explained and analyzed
- Network visualizations can provide intuitive pictures of complex social structures

Disadvantages

- Can be difficult to collect network data, especially for sociocentric networks
- Missing data in sociocentric networks can be very problematic
- Certain kinds of research questions—especially those comparing certain structural features of two networks or looking at changes in networks over time—may require advanced statistical models for social networks that are not widely available or easy to implement.

## Focus Groups

Focus groups combine elements of both interviewing and participant observation. The focus group session is, indeed, an interview—not a discussion group, problem-solving session, or decision-making group. At the same time, focus groups capitalize on group dynamics. The hallmark of focus groups is the explicit use of the group interaction to generate data and insights that would be unlikely to emerge otherwise. The technique inherently allows observation of group dynamics, discussion, and firsthand insights into the respondents' behaviors, attitudes, language, etc.

Focus groups are a gathering of eight to 12 people who share some characteristics relevant to the evaluation. Originally used as a market research tool to investigate the appeal of various products, the focus group technique has been adopted by other fields, such as education, as a tool for data gathering on a given topic. Initially, focus groups took place in a special facility that included recording apparatus (audio and/or visual) and an attached room with a one-way mirror for observation. There was an official recorder, who may or may not have been in the room. Participants were paid for attendance and provided with refreshments. As the focus group technique has been adopted by fields outside of marketing, some of these features, such as payment or refreshments, have sometimes been eliminated.

With the advent of new technologies, the focus group approach is taking new forms. In addition to telephone focus groups that permit a geographically dispersed "group" to be conducted, focus groups are also being conducted using web-based technologies.

### When to Use Focus Groups

Focus groups can be useful at both the formative and summative stages of an evaluation. They provide answers to the same types of questions as in-depth interviews, except that they take place in a social context. Specific applications of the focus group method in evaluations include:

- Identifying and defining problems in project implementation
- Pretesting topics or idea
- Identifying project strengths, weaknesses, and recommendations
- Assisting with interpretation of quantitative findings
- Obtaining perceptions of project outcomes and impacts
- Generating new ideas

Although focus groups and in-depth interviews share many characteristics, they should not be used interchangeably. Factors to consider when choosing between focus groups and in-depth interviews are displayed in Exhibit 16.

**Exhibit 16.—Which to use: Focus groups or in-depth interviews?**

| Factors to consider | Use focus groups when… | Use interviews when… |
|---|---|---|
| Group interaction | interaction of respondents may stimulate a richer response or new and valuable thought. | group interaction is likely to be limited or nonproductive. |
| Group/peer pressure | group/peer pressure will be valuable in challenging the thinking of respondents and illuminating conflicting opinions. | group/peer pressure would inhibit responses and cloud the meaning of results. |
| Sensitivity of subject matter | subject matter is not so sensitive that respondents will temper responses or withhold information. | subject matter is so sensitive that respondents would be unwilling to talk openly in a group. |
| Depth of individual responses | the topic is such that most respondents can say all that is relevant or all that they know in less than 10 minutes. | the topic is such that a greater depth of response per individual is desirable, as with complex subject matter and very knowledgeable respondents. |
| Data collector fatigue | it is desirable to have one individual conduct the data collection; a few groups will not create fatigue or boredom for one person. | it is possible to use numerous individuals on the project; one interviewer would become fatigued or bored conducting all interviews. |
| Extent of issues to be covered | the volume of issues to cover is not extensive. | a greater volume of issues must be covered. |
| Continuity of information | a single subject area is being examined in depth and strings of behaviors are less relevant. | it is necessary to understand how attitudes and behaviors link together on an individual basis. |
| Experimentation with interview guide | enough is known to establish a meaningful topic guide. | it may be necessary to develop the interview guide by altering it after each of the initial interviews. |
| Observation by stakeholders | it is desirable for stakeholders to hear what participants have to say. | stakeholders do not need to hear firsthand the opinions of participants. |
| Cost and training | quick turnaround is critical, and funds are limited. | quick turnaround is not critical, and budget will permit higher cost. |
| Availability of qualified staff | focus group facilitators need to be able to control and manage groups. | interviewers need to be supportive and skilled listeners. |

## Observations

Observational techniques are methods by which an individual or individuals gather firsthand data on the interventions, processes, or behaviors being studied. They provide evaluators with an opportunity to collect data on a wide range of behaviors, to capture a great variety of interactions, and to openly explore the evaluation topic. By directly observing operations and activities, the evaluator can develop a holistic perspective, i.e., an understanding of the context within which the project operates. This may be especially important where it is not the event that is of interest, but rather how that event may fit into, or be affected by, a sequence of events. Observational approaches also allow the evaluator to learn about issues the participants or staff may be unaware of or that they are unwilling or unable to discuss candidly in an interview or focus group.

### When to Use Observations

Observations can be useful during both the formative and summative phases of evaluation. For example, during the formative phase, observations can be useful in determining whether or not the project is being delivered and operated as planned. During the summative phase, observations can be used to determine whether or not the project has been successful. For example, the technique would be especially useful in directly examining teaching methods employed by the faculty in their own classes after program participation. Exhibit 17 shows the advantages and disadvantages of observations.

**Exhibit 17.—Advantages and disadvantages of observations**

**Advantages:**

- Provide direct information about behavior of individuals and groups
- Permit evaluator to enter into and understand situation/context
- Provide good opportunities for identifying unanticipated outcomes
- Exist in natural, unstructured, and flexible setting

**Disadvantages:**

- Expensive and time consuming
- Need well-qualified, highly trained observers; may need to be content experts
- May affect behavior of participants
- Selective perception of observer may distort data
- Behavior or set of behaviors observed may be atypical

## Tests

Tests provide a way to assess subjects' knowledge and capacity to apply this knowledge to new situations. Tests take many forms. They may require respondents to choose among alternatives (select a correct answer, select an incorrect answer, select the best answer), to cluster choices into like groups, to produce short answers, or to write extended responses. A question may address a single outcome of interest or lead to questions involving a number of outcome areas.

Tests provide information that is measured against a variety of standards. The most popular test has traditionally been norm-referenced assessment. Norm-referenced tests provide information on how the target performs against a reference group or normative population. In and of itself, such scores say nothing about how adequate the target's performance may be, only how that performance compares with the reference group. Other assessments are constructed to determine whether or not the target has attained mastery of a skill or knowledge area. These tests, called criterion-referenced assessments, provide data on whether important skills have been reached but say far less about a subject's standing relative to his/her peers. A variant on the criterion-referenced approach is proficiency testing. Like the criterion-referenced test, the proficiency test provides an assessment against a level of skill attainment, but it also includes standards for performance at varying levels of proficiency, typically a three- or four-point scale ranging from below basic to advanced performance. Today, most state testing programs use some kind of proficiency scores to described outcomes.

Criticisms of traditional, short-answer tests focus on the fragmented and superficial nature of these tests and the consequent, negative influence they have on instruction, especially where the tests are used for high-stakes decision making. Critics call instead for assessments that are more authentic in nature, involving higher order thinking skills and the coordination of a broad range of knowledge. Proposed alternatives require students to engage in solving more complex problems and may involve activities such as oral interviews, group problem-solving tasks, portfolios, or personal documentation. These alternatives have not proven to be feasible in large-scale assessment programs, but may be very useful in smaller scale research efforts.

**When to Use Tests**

Tests are used when one wants to gather information on the status of knowledge or the change in status of knowledge over time. They may be used purely descriptively or to determine whether the test taker qualifies in terms of some standard of performance. Changes in test performance are frequently used to determine whether a project has been successful in transmitting information in specific areas or influencing the thinking skills of participants. Exhibit 18 shows the advantages and disadvantages of tests.

In choosing a test, it is important to assess the extent to which the test measures knowledge, skills, or behaviors that are relevant to your program. Not all tests measure the same things, nor do they do so in the same ways. The critical word here is "alignment." There are a number of different ways to assess alignment. Some useful suggestions are offered at http://archive.wceruw.org/nise/.

**Exhibit 18.—Advantages and disadvantages of tests**

The advantages and disadvantage of tests depend largely on the type of test being considered and the personal opinion of the stakeholder. However, the following claims are made by proponents.

**Advantages:**

- Provide objective information on what the test taker knows and can do
- Can be constructed to match a given curriculum or set of skills
- Can be scored in a straightforward manner
- Are accepted by the public as a credible indicator of learning

**Disadvantages:**

- May be oversimplified and superficial
- May be very time consuming
- May be biased against some groups of test takers
- May be subject to corruption via coaching or cheating

## Other Methods

The last section of this chapter outlines less common, but potentially useful qualitative methods for project evaluation. These methods include document studies, key informants, and case studies.

### Document Studies

Existing records often provide insights into a setting and/or group of people that cannot be observed or noted in another way. This information can be found in document form. Lincoln and Guba (1985) defined a document as "any written or recorded material" not prepared for the purposes of the evaluation or at the request of the inquirer. Documents can be divided into two major categories: public records and personal documents (Guba and Lincoln, 1981).

Public records are materials created and kept for the purpose of "attesting to an event or providing an accounting" (Lincoln and Guba, 1985). Public records can be collected from outside (external) or within (internal) the setting in which the evaluation is taking place. Examples of external records are census and vital statistics reports, county office records, newspaper archives, and local business records that can assist an evaluator in gathering information about the larger community and relevant trends. Such materials can be helpful in better understanding the project participants and making comparisons among groups/ communities.

For the evaluation of educational innovations, internal records include documents such as student transcripts and records, historical accounts, institutional mission statements, annual reports, budgets, grade and standardized test reports, minutes of meetings, internal memoranda, policy manuals, institutional histories, college/university catalogs, faculty and student handbooks, official correspondence, demographic material, mass media reports and presentations, and descriptions of program development and evaluation. They are particularly useful in describing institutional characteristics, such as backgrounds and academic performance of students, and in identifying institutional strengths and weaknesses. They can help the evaluator understand the institution's resources, values, processes, priorities, and concerns. Furthermore, they provide a record or history that is not subject to recall bias.

Personal documents are first-person accounts of events and experiences. These "documents of life" include diaries, portfolios, photographs, artwork, schedules, scrapbooks, poetry, letters to the paper, etc. Personal documents can help the evaluator understand how the participant sees the world and what she or he wants to communicate to an audience. Unlike other sources of qualitative data, collecting data from documents is relatively invisible to, and requires minimal cooperation from, persons within the setting being studied (Fetterman, 1989). Information from documents also can be used to generate interview questions or identify events to be observed. Furthermore, existing records can be useful for making comparisons (e.g., comparing project participants to project applicants, project proposal to implementation records, or documentation of institutional policies and program descriptions prior to and following implementation of project interventions and activities).

The usefulness of existing sources varies depending on whether they are accessible and accurate. When using such instruments, it is advisable to do a quick scan to assess data quality before undertaking extensive analysis. Exhibit 19 shows the advantages and disadvantages of document studies.

**Exhibit 19.—Advantages and disadvantages of document studies**

**Advantages:**

- Available locally
- Inexpensive
- Grounded in setting and language in which they occur
- Useful for determining value, interest, positions, political climate, public attitudes
- Provide information on historical trends or sequences
- Provide opportunity for study of trends over time
- Unobtrusive

**Disadvantages:**

- May be incomplete
- May be inaccurate or of questionable authenticity
- Locating suitable documents may pose challenges
- Analysis may be time consuming and access may be difficult

**Key Informant**

A key informant is a person (or group of persons) who has unique skills or professional background related to the issue/intervention being evaluated, is knowledgeable about the project participants, or has access to other information of interest to the evaluator. A key informant can also be someone who has a way of communicating that represents or captures the essence of what the participants say and do. Key informants can help the evaluation team better understand the issue being evaluated, as well as what the project participants say and do. They can provide important contextual information on the current implementation environment, as well as relevant historical background. Key informants can be surveyed or interviewed individually or through focus groups.

Many different types of people can play the key informant role. At a university, a key informant could be a dean, a grants officer, or an outreach coordinator. In a school system, key informants range from a principal, to the head of a student interest group, to a school board member. Both the context and the politics of a situation affect who may be seen in the key informant role.

The use of advisory committees is another way of gathering information from key informants. Advisory groups are called together for a variety of purposes:

- To represent the ideas and attitudes of a community, group, or organization
- To promote legitimacy for the project
- To advise and recommend
- To carry out a specific task

Members of such a group may be specifically selected or invited to participate because of their unique skills or professional background; they may volunteer; they may be nominated or elected; or they may come together through a combination of these processes. Exhibit 20 shows the advantages and disadvantages of key informants.

**Exhibit 20.—Advantages and disadvantages of using key informants**

**Advantages:**

- Information concerning causes, reasons, and/or best approaches is gathered from an "insider" point of view
- Advice/feedback increases credibility of study pipeline to pivotal groups
- May have side benefit to solidify relationships among evaluators, clients, participants, and other stakeholders

**Disadvantages:**

- Time required to select and get commitment may be substantial
- Relationship between evaluator and informants may influence type of data obtained
- Informants may interject own biases and impressions
- Disagreements among individuals may be hard to resolve

**Case Studies**

Classical case studies depend on ethnographic and participant observer methods. They are largely descriptive examinations, usually of a small number of sites (small towns, projects, individuals, schools) where the evaluator is immersed in the life of the site or institution, combs available documents, holds formal and informal conversations with informants, observes ongoing activities, and develops an analysis of both individual and cross-case findings.

Case studies can provide very engaging, rich explorations of a project or application as it develops in a real-world setting. Project evaluators must be aware, however, that doing even relatively modest, illustrative case studies is a complex task that cannot be accomplished through occasional, brief site visits. Demands with regard to design, data collection, and reporting can be substantial (Yin, 2002). Exhibit 21 shows the advantages and disadvantages of case studies.

**Exhibit 21.—Advantages and disadvantages of using case studies**

**Advantages:**

- Provide a rich picture of what is happening, as seen through the eyes of many individuals
- Allow a thorough exploration of interactions between treatment and contextual factors
- Can help explain changes or facilitating factors that might otherwise not emerge from the data

**Disadvantages:**

- Require a sophisticated and well-trained data collection and reporting team
- Can be costly in terms of the demands on time and resources
- Individual cases may be overinterpreted or overgeneralized

## Summary

There are many different types of data collection methods that can be used in any evaluation. Each has its advantages and disadvantages and must be chosen in light of the particular questions, timeframe, and resources that characterize the evaluation task. While some evaluators have strong preferences for quantitative or qualitative techniques, today

the prevailing wisdom is that no one approach is always best, and a carefully selected mixture likely provides the most useful information.

**References**


Durland, M.M., and Fredericks, K.A. (eds.). (2006) Social Network Analysis in Program Evaluation. *New Directions for Evaluation*, No. 7. San Francisco, CA: Jossey-Bass.

Fetterman, D.M. (1989). *Ethnography: Step by Step*. Applied Social Research Methods Series, Vol. 17. Newbury Park, CA: Sage.

Guba, E.G., and Lincoln, Y.S. (1981). *Effective Evaluation*. San Francisco, CA: Jossey-Bass.

Lincoln, Y.S., and Guba, E.G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

Patton, M.Q. (2001). *Qualitative Evaluation and Research Method*. 3rd Ed. Newbury Park, CA: Sage.

Yin, R.K. (2002). *Case Study Research.* 3rd Ed. Thousand Oaks, CA: Sage.

Chapter 7

# A GUIDE TO CONDUCTING CULTURALLY RESPONSIVE EVALUATIONS

Henry T. Frierson, Stafford Hood, Gerunda B. Hughes, and Veronica G. Thomas

Since the last edition of this Handbook and the initial chapter by Frierson, Hood, and Hughes (2002) that addressed the issue of conducting cultural responsive evaluation, there has been considerably more emphasis on culture, context, pluralism, and inclusiveness in project evaluation (e.g., Botcheva, Shih, and Huffman, 2009; Guzman, 2003; Hood, Hopson, and Frierson, 2005; Mertens, 2003; Thomas and Stevens, 2004; Thompson-Robinson, Hopson, and SenGupta, 2004; Zulli and Frierson, 2004; Hood, 2009; Hopson, 2009). Central in much of this discourse is the position that project conceptualization, design, implementation, and evaluation take place within a variety of historical, social, cultural, political, and economic contexts and that evaluation must take these myriad contexts into consideration. Understanding the influence of culture, particularly when evaluating projects serving diverse populations, is critical for strengthening the validity and utility of evaluation findings and for improving evaluation practice in accordance with the American Evaluation Association's (AEA) *Guiding Principles for Evaluators* and the Joint Committee on Standards for Educational Evaluation's *Program Evaluation Standards*. As such, there is growing recognition that cultural issues cannot be simply viewed as "error noise," but rather as part of what will inform the whole story of an evaluation, thereby filling in those "missing bricks" (Jolly, 2002) of foundational knowledge. With increased attention to the cultural context of evaluation, there are encouraging signs that evaluation practice is becoming more responsive to working in culturally diverse settings.

*Evaluation is based on an examination of impacts through lenses in which the culture of the participants is considered an important factor.*

This chapter discusses the importance of employing a culturally responsive approach when evaluating projects serving populations within cultural contexts unfamiliar to the project evaluator or projects involving individuals with cultural backgrounds different than that of the project evaluator. It examines cultural responsiveness at each of the critical phases of the evaluation process, showing how strategies commensurate with this approach can be applied to enhance the actual quality and utility of project evaluations. This updated version provides more illustrative examples of culturally responsive strategies used in project evaluations where cultural diversity is acknowledged and taken into account. Additionally, a new section on ethical considerations and cultural responsiveness has been added at the end of the chapter.

Culture is a cumulative body of learned and shared behavior, values, customs, and beliefs common to a particular group or society. In essence, culture is a predominant force shaping who we are. In doing project

evaluation, it is important to consider the cultural context in which the project operates and be responsive to it. How can an evaluation be culturally responsive? An evaluation is culturally responsive if it fully takes into account the culture of the program that is being evaluated. In other words, the evaluation is based on an examination of impacts through lenses in which the culture of the participants is considered an important factor, thus rejecting the notion that assessments must be objective and culture-free if they are to be unbiased. Moreover, a culturally responsive evaluation attempts to fully describe and explain the context of the program or project being evaluated.

Culturally responsive evaluators honor the cultural context in which an evaluation takes place by bringing needed, shared life experiences and understandings to the evaluation tasks at hand and hearing diverse voices and perspectives. This approach requires that evaluators critically examine culturally relevant but often neglected variables in program design and evaluation. In order to accomplish this task, the evaluator must have a keen awareness of the context in which the project is taking place and an understanding of how this context might influence the behavior of individuals in the project. Here, context denotes a broader concept that entails the combination of factors (including culture) accompanying the implementation and evaluation of a project that might influence its results (Thomas, 2004). Examples of these factors include geographic location, timing, political and social climate, economic conditions, and other things going on at the same time as the project. In other words, context is the totality of the environment in which the project takes place.

Why should a project director be concerned with the cultural context of a project undergoing evaluation? Simply put, as American society becomes increasingly diverse racially, ethnically, and linguistically, it is important that project designers, implementers, and evaluators understand the cultural contexts in which these projects operate. To ignore the reality of the influence of culture and to be unresponsive to the needs of the target population puts the program in danger of being ineffective and the evaluation in danger of being seriously flawed. Evaluation should serve the public good by presenting valid information about programs that have been properly evaluated.

*Cultural responsiveness is gaining recognition as a critical feature of the evaluation process.*

Being sensitive and responsive to the culture of the participants and the cultural environment in which the programs exists should be an important component of project evaluation. Fortunately, cultural responsiveness as it relates to evaluation is gaining recognition as a critical feature of the evaluation process, particularly for programs in which the participants' culture is acknowledged to have a major impact on project outcomes.

The benefits related to cultural responsiveness in evaluations are discussed in the literature. For example, LaFrance (2004) maintains that learning about and understanding tribal culture when conducting

evaluations in Indian communities result in evaluations that are more responsive to tribal programs and broad enough to accommodate and value different ways of knowing that are not typical in Western evaluation models. Thomas and LaPoint (2004/2005) and LaPoint and Jackson (2004) point to how co-constructing family involvement and placing an authentic emphasis on cultural and contextual relevance improved an evaluation of an urban family-school-community partnership program serving predominately African American populations. Using examples from projects serving Latino populations, Guzman (2003) stressed how consideration of cultural norms might lead to different and more accurate interpretation of evaluation findings.

*There are no culture-free evaluators, educational tests, or societal laws.*

## The Need for Culturally Responsive Evaluation

It may seem obvious to some, if not to most, professionals that cultural responsiveness should be an integral part of the project development and evaluation process. After all, who could argue against taking into account the cultural context when designing and conducting an evaluation? Doesn't everyone consider the cultural context? The answers to these questions are, respectively, "many" and "no." Apparently, not everyone agrees that implementing culturally responsive evaluation is a good idea. Essentially, there are two frequently stated arguments against using culturally responsive strategies and techniques in educational evaluations. First, there is the claim that evaluations should be culture-free. Second, some individuals argue that while an evaluation should take into account the culture and values of the project or program it is examining, it should not, however, be *responsive* to them.

Let us examine the first argument. Just as surely as there are no culture-free evaluations, there are no culture-free evaluators, educational tests, or societal laws. Our values are reflected in our social activities, whether they are educational, governmental, or legal. An evaluator's values, beliefs, and prejudices, in particular, can and do influence a number of critical aspects of the evaluation process. Thomas and McKie (2006) delineated seven ways in which this can occur, including influencing (a) what questions an evaluator asks and ultimately does not ask, (b) what an evaluator illuminates and ultimately minimizes, (c) what evaluation approach is used and ultimately not used, (d) what data are collected and ultimately overlooked, (e) how interpretations are made and whose interpretations are held in high or low esteem, (f) what conclusions are drawn and what conclusions are not considered, and (g) how results are presented and to whom such results are disseminated. The responsibility that educational evaluators have is to recognize their own personal cultural preferences and to make a conscious effort to minimize any undue influence they might have on the work.

The second argument, that educational evaluations should not be *responding* to the cultural contexts in which they are undertaken, is more troublesome. It is one thing to accept or recognize the reasonableness of

the requirement to describe the cultural context. It is quite another to adopt evaluation strategies that are consonant with the cultural context(s) under examination. It is precisely this last point of view that is being advocated in this chapter. Since the 1960s, the field of educational evaluation has come to recognize the role that fullness of description plays in a comprehensive evaluation process (e.g., Stake, 1967). In fact, it is becoming increasingly apparent that a responsive evaluation can greatly benefit the project and its stakeholders. Still, it remains all too rare that educational evaluation is designed to be responsive to the cultural context associated with the program or project that is being evaluated.

Culturally responsive evaluation does not consist of a distinct set of steps apart from any high-quality evaluation approach. Rather, it represents a holistic framework for thinking about and conducting evaluations in a culturally responsive manner. It is a process entailing the manner in which the evaluator plans the evaluation, engages the evaluand and its stakeholders, and takes into account the cultural and social milieu surrounding the program and its participants. Indeed, evaluation products can be greatly enhanced through use of this approach.

Culturally responsive evaluation legitimizes culturally specific knowledge and ways of knowing. For example, the NSF supported two grants to the American Indian Higher Education Consortium (AIHEC, 2009) to develop evaluation processes that accomplish three purposes: (a) being robust enough to accommodate and value different "ways of knowing" within indigenous epistemologies, (b) building ownership and a sense of community within groups of Indian educators, and (c) contributing efficiently to the development of high-quality and sustainable STEM education programs. In doing so, this project broadens the national evaluation discourse through the inclusion of indigenous epistemologies that are not typically included in Western evaluation models and currently serves as a model for the design and evaluation of culturally responsive educational interventions in tribal communities. In developing the culturally responsive indigenous evaluation framework, the project was guided by six principles: (a) tribal people have always had ways of assessing merit or worth based on traditional values and cultural expressions, and this knowledge should inform how evaluation is done in tribal communities; (b) evaluation should respect and serve tribal goals for self-determination and sovereignty; (c) an indigenous framing for evaluation should incorporate broadly held values while also remaining flexible and responsive to local traditions and cultural expressions; (d) evaluation is defined (i.e., its meaning, practice, and usefulness) in tribal community terms, and the community takes ownership of this process and does not merely respond to the requirements imposed by outsiders; (e) evaluators should use practices and methods from the field of evaluation that fit tribal communities' needs and circumstances; and (f) evaluation is an opportunity for learning from the tribal communities' own programs and work, as well as using what is learned to create strong, viable tribal communities. These AIHEC projects are good illustrative examples of work that resulted in

the articulation of tangible strategies for respecting and being responsive to cultural norms, beliefs, values, and behavior patterns across the entire evaluation process.

## Preparing for the Evaluation

At the start of the evaluation process, evaluators must carefully analyze the project's cultural and sociopolitical context as it presently exists to help establish the parameters of the evaluation effort. In preparation for the evaluation, collection of background data on the evaluand, including information on the cultural context of the project and its participants, is crucial. This information can be gathered through multiple venues such as informant interviews with directors of organizations and leaders of the community, ongoing group discussions with other key stakeholders, community forum, and feedback sessions with community members. It should be noted that in many culturally-based communities, the real leaders are not necessarily those individuals in appointed positions of power; instead, they may be the role models, information sources, and problem solvers within the community who do not hold any formal position of authority.

Communication and relational styles can vary tremendously between and within different ethnic and culturally-based populations, and these differences should be explored during the preparation phase to better plan and implement the evaluation. Unintended insensitivity to different and unfamiliar cultural norms can hamper communications and understandings and negatively affect accurate data collection. For example, it has been pointed out that the meaning of "silence" often varies across cultural groups, particularly when the evaluator and persons under study do not share similar positions of power, status, and privilege. In contexts of unequal power relationships, silence may be used by the less powerful persons to maintain control, dignity, and self-respect. Evaluators should be briefed on the cultural nuances of communication and relational styles of the cultural groups under consideration prior to the start of the evaluation. In other words, planning for the evaluation in culturally diverse settings involves preparing to accomplish the technical aspects of completing an evaluation as well as concerted emphasis on building relationships, establishing trust, and gaining an understanding of cultural styles and norms that might influence the behavior of people in programs.

Before the evaluation begins, there should be consensus on the purpose and goals of the evaluation between the evaluator and the project staff. Culturally responsive evaluators go beyond simply attending to the funder's agenda in evaluation to also hearing and infusing, to the extent feasible, the perspectives of the target community in determining the evaluation's purpose. The evaluators can substantially enhance their success in this effort through engaging key stakeholders, gaining their trust and cooperation, and facilitating their ownership of the evaluation.

*Multi-ethnic evaluation teams increase the chances of really hearing the voices of underrepresented students.*

Preparing for the actual evaluation and assembling an evaluation team is, of course, a critical stage in the evaluation process. At the outset, the sociocultural context in which the programs or projects are based must be taken into account. Situations where programs involve ethnically diverse participants and stakeholders call for the "creation of multi-ethnic evaluation teams to increase the chances of really hearing the voices of underrepresented students" (Stevens, 2000). Stevens reminds us that evaluators may, and often do, *listen* to what stakeholders say when they collect data on-site from students, teachers, parents, and other participants or stakeholders. But the crucial question she asks is: do they *hear* what those individuals are saying? Stevens implies that the evaluator or evaluation team must have the "shared lived" experience to truly hear what is being said. There are instances in evaluation practice supporting this argument. For example, a group of African American evaluators working in predominately African American urban school settings maintained that because of the shared racial and ethnic background they had with project staff and participants, the evaluator team was keenly aware of and sensitive to many of the contextual and cultural issues relevant to the lives of the children and family member being served (Thomas, 2004). They brought a different set of experiences to the urban school context than non-African American evaluators, which the evaluators argued increased their ability to engage stakeholders and better understand the verbal and nonverbal behaviors of the individuals being served.

In reality, it may not be practical to select members of the evaluation team who have the shared lived experiences of various racial or ethnic groups represented among participants in the project under study given the relatively small proportion of evaluators of color in the field. Thus, it is essential that members of the evaluation team *acquire* a fundamental understanding of the cultural norms and experiences of the individuals under consideration. The team can engage individuals familiar with the group being studied as informants, interpreters, and critical friends to the evaluation. These individuals can act as cultural guides to the community through translating cultural cues and advising the evaluation team on the cultural appropriateness of their evaluation approach. Hiring and training individuals from the community to serve on the evaluation team in various capacities is another strategy for enhancing the evaluation team's sensitivity and awareness to the cultural realities of the community under study. Further, evaluators who are not familiar with the cultural groups being studied should engage in an ongoing process of self-reflection and reflective adaptation. Self-reflection provides opportunities for evaluators to become acutely aware of their own cultural values, assumptions, prejudices, and stereotypes and how these factors may affect their evaluation practice within the particular setting. Reflective adaptation is the ability to acknowledge one's biases, listen to other world views, and integrate these varying views and interests as they relate to evaluation design and implementation (Botcheva, Shih, and Huffman, 2009). At the very least, the evaluator or evaluation team

should be fully aware of and responsive to the participants' and stakeholders' cultures, particularly as they relate to and influence the program.

Given the important role of the evaluation team, care should be taken in selecting its members. Those members, whenever possible, should be individuals who understand or at least are clearly committed to being responsive to the cultural context in which the project is based. Project directors should not, however, assume that racial/ethnic congruence among the evaluation team, participants, and stakeholders equates to cultural congruence or competence that is essential for carrying out culturally responsive evaluations (Thomas, 2004).

## Engaging Stakeholders

*Stakeholders play a critical role in all evaluations, especially culturally responsive ones.*

Stakeholder involvement has long been an expectation of good evaluation practice. Stakeholder involvement and relationship building are particularly critical when conducting evaluations of projects serving diverse and, oftentimes, marginalized populations. When individuals in minority communities feel marginalized or powerless, issues of power relations, status, and social class differentials between evaluators and the target population can impede the stakeholder engagement process. These issues must be worked through very carefully. In culturally responsive evaluation, which is inherently participatory in nature, stakeholders must be engaged and encouraged to become active participants in the construction of knowledge about their lives and communities.

When designing an evaluation that seeks to be culturally responsive, considerable attention must be given to the identification of the stakeholders. Often, identified stakeholders include those who are most vocal, most visible, and easiest to work with throughout the evaluation process, but ignoring other relevant stakeholders might result in failing to capture critical contextual aspects of the project under study, which potentially can lead to inaccurate judgments and conclusions. Issues related to the identification and prioritization of relevant stakeholders and gaining access to and getting the cooperation of the multiple stakeholder groups are evaluation challenges that can be more meaningfully addressed through engaging and collaborating with members of the community.

Stakeholders play a critical role in all evaluations, especially culturally responsive ones, since they can provide sound advice from the beginning (framing questions) to the end (disseminating the evaluation results) of the evaluation process. It is important to develop a stakeholder group representative of the populations the project serves, ensuring that individuals from all sectors have the chance for input. Indeed, those in the least powerful positions can be the most affected by the results of an educational evaluation. Providing key stakeholders, especially those who

traditionally have had less powerful roles, with opportunities to have a voice can minimize problems related to unequal distribution of power and status differentials that can be problematic in evaluations of projects' minority populations. For example, in evaluations of urban school reform initiatives serving African American populations undertaken by the Howard University Center for Research on the Education of Students Placed at Risk (CRESPAR), stakeholders (e.g., students, parents, school staff) were given multiple opportunities to ask questions, critique evaluative efforts, and provide input in myriad ways (Thomas, 2004). Evaluators entered the context gently, respectfully, and with a willingness to listen and learn in order to obtain stakeholder buy-in and to plan and implement a better evaluation. During the meetings with key stakeholders, CRESPAR evaluators asked questions, listened to stakeholders' concerns, discussed issues, and recorded responses. These activities created a climate of trust and respect from the stakeholders once they realized that their input was genuinely wanted, valued, and, to the extent possible, incorporated into the evaluation activities. Stakeholders were engaged by allowing them input into framing evaluation questions, developing instruments, collecting data, interpreting findings, and using and disseminating the findings.

Engage stakeholders in the evaluation process by inviting them to serve on project advisory boards or steering committees. These committees or boards can provide input into decisions about evaluation planning, design, implementation, and dissemination. In particular, they can collaborate with the evaluation team in framing evaluation questions, reviewing and providing feedback on instruments, interpreting findings, and developing recommendations. These groups can also assist with communication between the evaluation team and key individuals or groups of the project under study.

Failure to identify and engage stakeholders from the community being studied can be problematic at various levels. For example, in evaluations of projects in tribal communities, not including tribal members in the planning, implementation, and dissemination of the evaluation results is viewed as a serious affront to those involved as evaluation participants, and it is thought to have the potential of invalidating the evaluation results.

In individual projects such as the Louis Stokes Alliance for Minority Participation and the Alliance for Graduate Education for the Professoriate (LSAMP), if participants' and stakeholders' perceptions and views are not taken into account from a cultural perspective, the evaluation may be flawed, particularly if qualitative methods are employed. Moreover, even if quantitative methods are the primary methodological format, the various "voices" should be heard in the interpretation and presentation of the results.

## Identifying the Purpose(s) and Intent of the Evaluation

Another important step is to ensure that there is a clear understanding of the evaluation's purpose and intent. Generally speaking, as stated earlier, comprehensive project evaluation is designed to answer two basic questions: (a) Is the project being conducted as planned and is progress being made toward meeting its goals? (b) Ultimately, how successful is the project in reaching its goals? To answer these questions, three basic types of evaluations are conducted: process, progress, and summative. The first two types of evaluations are called formative evaluations because they assess and describe project operations in order to inform project staff (and stakeholders) about the status of the project. Summative evaluations, on the other hand, reveal whether and to what extent the project achieved its goals and objectives.

*Culturally responsive progress evaluations examine connections through culturally sensitive lenses.*

Process evaluations examine connections among project activities. Culturally responsive process evaluations examine those connections through culturally sensitive lenses. For example, the extent to which the project's philosophy compares and interacts with the cultural values of the target population and the extent to which effective cultural competence training is available for staff are two project activities that might be subjected to a process evaluation for evidence of cultural responsiveness. Careful documentation of the implementation of project activities is critical to making sense of the subsequent summative evaluation results. Having an evaluator or a team of evaluators that is culturally sensitive to the project environment will ensure that cultural nuances—large and small—will be captured and used for interpreting progress and summative evaluations.

Progress evaluations seek to determine whether the participants are progressing toward achieving the stated goals and objectives. Culturally responsive progress evaluations help determine whether the original goals and objectives are appropriate for the target population. In seeking to ascertain whether the participants are moving toward the expected outcomes, a culturally responsive progress evaluation can reveal the likelihood that the goals will be met, exceeded, or not exceeded given the project timeline and the results of the process evaluation.

Summative evaluations provide information about project effectiveness. Culturally responsive summative evaluations examine the direct effects of the project's implementation on the participants and attempt to explain the results within the context of the project and the lived experiences of the participants beyond the project. For example, improved student achievement is influenced by and correlated with a variety of school and personnel background variables. Thus, to fully measure the effectiveness of the project and determine its true rather than superficial worth, it is important to identify the correlates of participant outcomes (e.g., student achievement, student attitudes) and measure their effects as well.

## Framing the Right Questions

An important key to successful evaluation is to ensure that the proper and appropriate evaluation questions have been framed. For an evaluation to be culturally responsive, it is critical that the questions of the primary stakeholders have been heard and, where appropriate, addressed.

*It is critical that the questions of significant stakeholders have been heard and, where appropriate, addressed.*

The questions that will guide an educational evaluation are crucial to the undertaking and ultimately to the success of the venture. Poorly framed questions rarely yield useful information. Further, framing evaluative questions is *not* easily accomplished. In a culturally responsive evaluation, it is expected that the questions are carefully considered not only by the evaluator and project staff, but also by other stakeholders as well. It takes time and diligence to reach agreement on the questions to be pursued. One stakeholder group may care little about questions that are seen as vital by another group. However, it is crucial that all significant voices are heard.

Once a list of questions has been vetted to the satisfaction of the evaluation team and stakeholders, an important epistemological task and next step is to decide what type of evidence must be collected to answer the evaluative questions. This is not an easy task, but it should be undertaken *before* embarking on a culturally responsive evaluation. It avoids subsequent rejection of evidence by a stakeholder who might say, for example, "This is interesting, but it really isn't hard data." Stakeholders often will be interested in the results that bear on one group over all others. If one particular group has not been involved or asked questions they consider as key, then the rest of the data may be viewed as suspect or irrelevant.

*Questions regarding what constitutes acceptable evidence should be discussed before conducting the evaluation.*

Discussions of what is important, and how we will know if we have acceptable evidence, are often messy and may be heated. The discussions, however, are always *necessary*. A more democratic approach to evaluation increases the need for competent evaluators who have a shared lived experience with the stakeholders (Hood, 2000). A democratic process also increases the likelihood that evaluative efforts will have all voices represented.

In a culturally responsive evaluation approach, the evaluators must be reflective, that is, have an awareness of their contributions to the construction of meaning throughout the evaluation process and acknowledge the impossibility of remaining totally detached from the topic under study. Even after questions are identified, it would be helpful if an evaluator asked him or herself three important questions before moving forward: (a) Does the way in which the evaluation questions are defined limit what can be found? (b) Can the evaluation questions be studied differently than initially articulated? (c) How might different ways of studying the evaluation questions give rise to a different and, potentially, more expanded understanding of the phenomenon under

study? Answering these questions through a cultural lens, and making appropriate modifications to the originally framed questions, can result in a more responsive evaluation. Additionally, a fourth and very important question must be addressed regardless of the evaluation approach: Can the evaluation questions actually be answered based on the data sources available?

## Designing the Evaluation

After the evaluation questions have been properly framed, sources of data have been identified, and the type of evidence to be collected has been decided, it is then time to identify the appropriate evaluation design.

*The need to train data collectors in evaluation studies is great.*

A good design offers a unique opportunity to maximize the quality of the evaluation. A culturally responsive evaluation approach does not consider a particular design as correct or universally applicable. In fact, there are a number of different evaluation designs (e.g., quasi-experimental, experimental, ethnographic, case study) that can be used to organize the processes of data collection and analysis and subsequently answer the evaluation questions. The evaluation design that one uses does not necessarily need to be elaborate. It just needs to be appropriate for what the evaluator wants to do for an effective evaluation study.

Most comprehensive evaluation designs are mixed-methods, that is, they have both qualitative and quantitative components in a single study in an effort to increase the scope of confidence in the findings. Each component provides data in a format that is different from the other but can be complementary. Increasingly, evaluations are relying on mixed-methods, recognizing that both approaches are valuable and have something unique to offer. Mixed-methods might be especially relied upon in culturally responsive evaluations as a way of gathering information that more fully addresses the complexities in culturally diverse settings. This approach should provide a better opportunity for documenting the complexities of processes, progress, and outcomes occurring in culturally complex and diverse settings. For example, an evaluator examining student achievement might decide to look at quantitative outcomes such as students' grades based upon teacher-developed tests, textbook tests, or standardized test scores; in addition, the evaluator may also look at various qualitative indicators such teacher-student interactions, student-student interactions, how students are taught, teacher qualitative reports of students, and school culture and environment. While quantitative data might demonstrate differences among subgroups of students, qualitative data would be particularly useful in gathering more nuanced information on the factors likely contributing to these differences.

Designs that incorporate data collection at multiple times provide an opportunity to examine the degree to which some aspect of the participants' behavior changed as a result of the project intervention(s). On the other hand, when comparison or control groups can be

incorporated into pretest/posttest designs, evaluators may be able to determine the extent to which some aspect of participants' behavior changed relative to where it would have been had they not been subject to the project intervention(s). It should be noted, however, that a culturally responsive evaluation approach does not advocate a particular approach toward inquiry and does not reside in either a quantitative or qualitative camp. Value in both approaches is seen by these authors. The view here, however, is that the perspective the evaluator brings to the evaluation task is the key entity.

## Selecting and Adapting Instrumentation

*Previous use does not guarantee cultural responsiveness.*

Instrumentation provides the means for collecting much of the data for program and project evaluation. Therefore, it is very important that instruments be identified, developed, or adapted to reliably capture the kind and type of information needed to answer the evaluation questions. Also at issue is the validity of the inferences about the target population that are drawn from data collected using evaluation instruments. It is preferable to use instruments that have some history. Piloting instruments to accumulate substantive evidence of validity and reliability is critically important. Yet, the previous use of standard methods to accumulate evidence of validity and reliability does not guarantee cultural responsiveness. Oftentimes, measures that have been normed on a cultural group different from the target population are used in the evaluation process. In such instances, additional pilot testing of the instruments should be done with the cultural group or groups involved in the study to examine their appropriateness. If problems are identified, refinements and adaptations of the instruments should be made so that they are culturally sensitive and thus provide reliable and valid information about the target population.

Given the growing number of projects serving Latino and other populations in which English was not their first language, evaluations are increasingly faced with the need to employ instruments in the primary language of the clients under study. As a result, instrument translation is becoming an integral part of the instrument-development process. Obviously, poor translation of evaluation instruments can be a serious problem. It can, in essence, render the data collected from such an instrument as valueless. There are various strategies for instrument translation. One common strategy has been to provide the English version of an instrument to a native speaker of the target population and ask that person to translate the instrument into the target language. This method, referred to as "simple direct translation," is inadequate.

It is ideal to use one of two widely used translation methods in national and international studies to better ensure accuracy of the translation. These include forward/backward translation (FBT) or translation by committee (TBC). The FBT technique, which is the generally preferred method of translation, involves having one individual translate an

instrument (document A) into another language (document B) and another person translating the resulting instrument back to the original language (document C). If documents A and C are determined to be equivalent, then document B is assumed to be a good translation (Marin and Marin, 1991). A caution noted with the FBT approach is that if both the forward and backward translators share common misconceptions about the target language and its semantic shadings, they could easily make similar mistakes in both translating and backward translating. Also, if the forward translator is excellent but the backward translator is not, the resultant outcome may be less than desirable. The TBC method, often more suitable to complete a transition in a short timeframe, involves using a bilingual panel to translate the instrument into the desired language. A more recent translation approach is multiple forward translations (MFT), whereby translators create two or more forward translations, which are then reconciled by another independent translator. MFT generally can be done more quickly than FBT because the two forward translations can be done concurrently. With any translation method, it is essential that all translators or committee members involved be bilingual, bicultural, and familiar with the target population. In assessing the validity of the translated instruments, it is recommended that at a minimum, the evaluator seek semantic and content equivalence. Semantic equivalence refers to the agreement between different language versions of the instruments. Content equivalence ensures that each item's content is relevant in each culture.

## Collecting the Data

As noted earlier, culturally responsive evaluation can make substantial use of qualitative evaluation techniques. Data collected through observations, interviews, and focus groups can be crucial for capturing rich information on the cultural contexts of the project and/or community under study. An important aspect of qualitative methodology is allowing participants to "voice" their own reality. Storytelling, chronicles, parables, poetry, observations, interviews, focus groups, and revisionist histories are all legitimate forms of data collection for knowledge generation and giving voice to participants. Storytelling, for example, has been used as a method for collecting rich cultural, historical, and other contextual information that may ultimately explain the behavior of people in projects under study. Additionally, in culturally responsive evaluation, use of a qualitative methodology often yields information that allows the evaluation team to select, adapt, or develop quantitative instruments to better capture the environment under consideration.

One of the tenets of qualitative methodology is that the individual who collects the data is also the *instrument*. Another tenet of qualitative methodology, as well as quantitative methodology, is that a poorly designed or improperly prepared instrument provides invalid data. Consequently, when collecting qualitative data directly from individuals, e.g., via interviews or observations, if those who are collecting and recording the data are not attuned to the cultural context in which the project is situated, the collected data could be invalid. While it may not appear to matter very much whether a person collecting student test papers in the classrooms is culturally responsive, cultural responsiveness does matter in many forms of data collection. In truth, it may indeed matter *how* the test papers are handed out to the students, *how* the test is introduced, and *what* the atmosphere is at the site where the students are being tested. The situation becomes far more complex in the collection of evaluative information through observations and interviews. The need to train data collectors in evaluation studies is great and, unfortunately, largely overlooked. Training them to understand the culture in which they are working is an even rarer event.

*Too often the nonverbal behaviors are treated as "error variance" in the observation and ignored.*

There may not be much an evaluation team can do about the age, gender, race, and appearance of its members, but to deny that such factors influence the amount and quality of the data is imprudent. One thing that can be done to increase the probability of gathering evaluative information in a culturally responsive manner is for the project director to ensure that the principal evaluator and team members involved in the data collection know what they are hearing and observing.

Nonverbal behaviors can often provide a key to data interpretation among culturally diverse populations. One African American psychologist, Naim Akbar (1975 as cited in Hale-Benson, 1982), describes a few nonverbal behaviors in African American children. He notes that the African American child "expresses herself or himself through considerable body language, adopts a systematic use of nuances of intonation and body language, such as eye movement and position, and is highly sensitive to others' nonverbal cues of communication." When observing African Americans participating in the project under evaluation, much could be lost toward reaching "understanding." Too often the nonverbal behaviors are treated as "error variance" in the observation and ignored. The same can be true when interviewing an African American program participant and stakeholder. In one sense, the evaluators have to know the territory. For example, Floraline Stevens (2000) described how she and her colleagues overcame difficulties attendant to being responsive to culture during an evaluation project within a large metropolitan school district. She pointed out that their extensive knowledge of the culture in the classroom and cultural background of the students overcame difficulties in collecting accurate data.

Lack of knowledge about cultural context is quickly evident when interview data are examined. Reviews of interview transcripts and

*Disaggregation of collected data is a procedure that warrants increased attention.*

observation protocol data that are done by reviewers without the ability to interpret meaning based on the (largely) unwritten rules of cultural discourse are likely to result in interpretations that are more frequently wrong than right (Smith, 1999; Nelson-Barber et al., 2005). Similarly, subsequent discussions of flawed reviews limit communication and ultimately doom the possibility of shared understanding between participants and stakeholders of color and the evaluator who proves to be culturally nonresponsive.

Knowledgeable trainers who use the medium of videotaping can and have produced considerable improvement in the skills of interviewers who must collect data in cultural settings unfamiliar to them. The training process can be very revealing for participants who seek to understand more about the nonverbal language they communicate and their own flawed communication habits. If interviewer training is entered with a spirit of openness and self-improvement, collection of culturally responsive evaluative data is greatly facilitated. Similar improvements in data collection and interpretation through observation can be achieved through intensive training and mentoring. Although the authors commend such training, inservice training is not the preferred solution. Greater and longer lasting improvements in the collection of culturally responsive evaluative data and the conduct of project evaluations can be realized principally by recruiting evaluation data collectors and analysts who already possess a shared lived experience with those who are being evaluated.

## Analyzing the Data

It is possible, though possibly shortsighted, to conduct statistical analyses and examine test score distributions without much concern for the cultural context in which the data were collected. Rather, it is both desirable and prudent that the analysis of interview data and the interpretation of descriptions of behavior related to projects undergoing evaluation be achieved with considerable sensitivity to, and understanding of, the cultural context in which the data are gathered. Determining an accurate meaning of what has been observed is central in culturally responsive evaluation. Having adequate understanding of cultural context when conducting an evaluation is important, but the involvement of evaluators who share a lived experience may be even more essential. The charge for minority evaluators is to go beyond the obvious.

Knowing the language of a group's culture guides one's attention to the nuances in how language is expressed and the meaning it may hold beyond the mere words. The analyst of data gathered in a culturally diverse context may serve as an interpreter for evaluators who do not share a lived experience with the group being evaluated.

To this end, a good strategy is to organize review panels principally comprising representatives from stakeholder groups. The review panels examine the findings gathered by the principal evaluator and/or an evaluation team. When stakeholder groups review evaluative findings, they may discover that views of the evaluators concerning the meaning of evaluative data might not necessarily be aligned with those of the review panel. The results of the deliberations of review panels will not lend themselves necessarily to simple, easy answers, but they will more accurately reflect the complexity of the cultural context in which the data were gathered and lead toward more accurate interpretations.

Data analyses from a culturally responsive approach seek to better understand how contextual conditions affect outcomes of people in projects. Culturally responsive evaluations use multiple strategies to analyze quantitative data to reveal a more complete picture of what is occurring within the environment under study. Disaggregation of collected data is a procedure that has gained increased attention in education and the social sciences in general, particularly since the passage of the No Child Left Behind Act of 2001, which requires students' standardized test scores to be disaggregated by economic background, race and ethnicity, English proficiency, and disability. Disaggregation is a method of "slicing" the data in order to examine the distribution of important variables for different subgroups in the population under study. For example, an evaluator may find that the average score on the nation's eighth grade science test for all eighth graders is 149; however, the average score on the same test for a particular ethnic subgroup of eight graders is likely to be substantially higher (or lower) than the average for all eighth graders.

Disaggregation of data sets is highly recommended because evaluative findings that dwell exclusively on whole-group statistics can blur rather than reveal important information. Even worse, they may be misleading. For example, studies that examine the correlates of *successful* minority students rather than focusing exclusively on the correlates of those who fail are important. It can be enlightening to scrutinize the context in which data that are regarded as outliers occur. The examination of a few successful students, in a setting that commonly produces failure, can be as instructive for project improvement as an examination of the correlates of failure for the majority.

Another data analysis procedure is to cross tabulate or, as it has been called, "dice" the data. Dicing the data involves a two-step process: first "slice" a simple statistic by race, socioeconomic status, or some other important cultural variable, then "dice" that statistic by another factor such as educational opportunity. It should be noted that disaggregating, or slicing, by a single racial or ethnic group may be insufficient. Since there is vast diversity within various ethnic groups, it is sometimes advisable to further disaggregate within group patterns. For example, an evaluator could analyze data for a Latino student population by recent immigrant status vs. second- or third-generation status. Similarly, it may be less valuable to lump all Black student ethnic groups together, but

instead more instructive to disaggregate by Black, American-born vs. immigrant from Africa and/or the West Indies.

In sum, it should be kept in mind that the data do not speak for themselves nor are they self-evident; rather, they are given voice by those who interpret them. The voices that are heard are not only those who are participating in the project, but also those of the analysts who are interpreting and presenting the data. Deriving meaning from data in project evaluations that are culturally responsive requires people who have some sense of the context in which the data were gathered.

## Disseminating and Using the Results

*Evaluation results should be viewed by audiences as not only useful, but truthful as well.*

Dissemination and utilization of evaluation outcomes are certainly important components in the overall evaluation process. One frequent outcome from dissemination of results of evaluations of programs serving culturally diverse minority communities is the tendency to attribute identified problems to the individuals or communities under study. A strategy that has been successfully utilized in culturally responsive evaluations in order to decrease potential backlash is to have individuals from the community review the findings before they are disseminated. From such a review, members of the community could provide cultural insights that help expand and enrich the interpretation of the evaluation findings. Also, there must be concerted efforts to close the "relevance gap" (Stanfield, 1999) between how much the evaluation data and their interpretations are congruent with the experiences of real people in the community under study. Moreover, a critical key is to conduct an evaluation in a manner that increases the likelihood that the results will be perceived as useful and, indeed, used. Culturally responsive evaluations can increase that likelihood. Hence, evaluation results should be viewed by audiences as not only useful, but truthful as well (Worthen, Sanders, and Fitzpatrick, 1997).

Information from good and useful evaluations should be widely disseminated. Further, communications pertaining to the evaluation process and results should be presented clearly so that they can be understood by all of the intended audiences.

Michael Q. Patton (1991) pointed out that evaluation should strive for accuracy, validity, and believability. Patton (2008) further stated that evaluation should assure that the information from it is received by the "right people." Building on his cogent observation, we would add that the "right people" are not restricted to the funding agency and project or program administration and staff, but should include a wide range of individuals who have an interest or stake in the program or project.

Culturally responsive evaluation encourages and supports using evaluation findings in ways to create a positive change in individuals' lives that might be affected by these findings. The dissemination and use

of evaluation outcomes should be thought through early when preparing an evaluation, that is, during the evaluation-planning phase. Moreover, the use of the evaluation should be firmly consistent with the actual purposes of the evaluation. Further, the purpose of the evaluation should be well defined and clear to those involved in the project itself.

As we talk about dissemination, our discussion comes full circle, and we return to the earliest steps in evaluation design, the evaluation questions. These questions themselves are always keys to a good evaluation—those that would provide information that stakeholders care about and on which sound decisions can be based must always guide the work. The right questions, combined with the right data collection techniques, can make the difference between an evaluation that is only designed to meet limited goals of compliance and one that meets the needs of the project and those who are stakeholders in it. Applying the principles of culturally responsive evaluation can enhance the likelihood that these ends will be met, and that the real benefits of the intervention can be documented.

## Ethical Considerations and Cultural Responsiveness

In evaluations, ethical decisions arise throughout the entire evaluation process, from conceptualization and design, data gathering, analysis and synthesis, data interpretation, and report writing to dissemination of findings. However, the evaluator is often faced with increased ethical responsibilities when conducting evaluations of projects serving culturally diverse populations. While some ethical considerations are quite obvious (e.g., doing no physical harm to participants), other ethical issues may be more subtle (e.g., the right of evaluators to impose their ideology on the people being studied, unequal power relations between the evaluator and those being observed or examined in the evaluation study, and the right of oppressed individuals to help shape evaluation questions and interpretations) (Thomas, 2009). As conceptualized, culturally responsive evaluation carefully takes into consideration these factors in an effort to conduct evaluation studies that are ethical and socially just.

Two types of ethics are particularly relevant in culturally responsive evaluations: (a) procedural ethics and (b) relational ethics. Procedural ethics, which are critical in all research, albeit evaluation is not research, involve those mandated by Institutional Review Boards (IRBs) to ensure that the study's procedures adequately deal with the ethical concerns of informed consent, confidentiality, right to privacy, freedom from deception, and protection of participants from harm. IRBs, however, generally give emphasis to assessing risks to individuals without paying attention to risks to communities, a condition that potentially has considerable ethical implications for evaluations focusing on marginalized communities (Minkler, 2004). Relational ethics recognize and value mutual respect, dignity, and the connectedness between the researcher and the researched and between the researchers and the communities in which they live and work (Ellis, 2007). Culturally

responsive evaluations pay special attention to risks to both individuals and communities, as well as remain mindful on building relationships of trust and mutual respect. Failure to do so might further marginalize the already distressed communities that projects are designed to serve. There are also ethical issues related to the evaluator's respect for local customs, values, and belief systems that should be taken into consideration in culturally diverse communities.

## Conclusions

With increasing recognition of the influence of culture on the attitudes and behaviors of individuals in education and other social programs, the time has come for cultural responsiveness to assume a central place in project evaluation practices. Within and across programs, diversity is often a critical feature that encompasses a variety of cultures (and subcultures), ethnicities, religions, languages, orientations, and values within the context of environmental and economic influences. Evaluators must seek authentic understanding of how a project functions within the context of diverse cultural groups to enhance confidence that they have asked the right questions, gathered correct information, drawn valid conclusions, and provided evaluation results that are both accurate and useful. Culturally responsive evaluations foreground issues of importance when working with any culturally-based groups such as attending to the influence of race, gender, ethnicity, class, and other factors that might be dismissed though they are central elements of individuals' lived experiences and realities. This approach to evaluation targets the environment as well as individuals operating within that environment through principles of stakeholder engagement, cooperation, collaboration, and the provision of data that a project and other relevant stakeholders can use to better understand a project's operations and outcomes within its cultural framework.

## References

American Evaluation Association. (2004, rev.). *Guiding Principles for Evaluators*. Fairhaven, MA: Author. (Full text available online at http://www.eval.org.)

American Indian Higher Education Association (AIHEC). (2009). *Indigenous Evaluation Framework: Telling Our Story in Our Place and Time.* Alexandria, VA: Author.

Botcheva, L., Shih, J., and Huffman, L. C. (2009). Emphasizing Cultural Competence in Evaluation. *American Journal of Evaluation,* 30(2): 176-188.

Ellis, C. (2007). Telling Secrets, Revealing Lives: Relational Ethics in Research With Intimate Others. *Qualitative Inquiry,* 13(1): 3-29.

Frierson, H. T ., Hood, S., and Hughes, G. B. (2002). Culturally Responsive Evaluation and Strategies for Addressing It. In *The User Friendly Evaluation Handbook.* Arlington, VA: National Science Foundation.

Gordon, E. W. (1998). Producing Knowledge and Pursuing Understanding: Reflections on a Career of Such Effort. AERA Invited Distinguished Lectureship. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, 13 April.

Guzman, B. L. (2003). Examining the Role of Cultural Competency in Program Evaluation: Visions for New Millennium Evaluators. In *Evaluating Social Programs and Problems: Visions for the New Millennium,* edited by S. I. Donaldson and M. Scriven, 167-181. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hale-Benson, J. (1982). *Black Children: Their Roots, Culture, and Learning Styles.* Rev. Ed. Baltimore: Johns Hopkins University Press.

Hopson, R. K. (2009). Reclaiming Knowledge at the Margins: Culturally Responsive Evaluation in the Current Evaluation Moment. In *The SAGE International Handbook of Educational Evaluation,* edited by K. Ryan and B. Cousins, 429-446. Thousand Oaks, CA: Sage Publications.

Hood, S. (2000). Commentary on Deliberative Democratic Evaluation. In *Evaluation as a Democratic Process: Promoting Inclusion, Dialogue, and Deliberation,* edited by K. Ryan and L. DeStefano. New Directions for Program Evaluation, No. 85. San Francisco: Jossey-Bass.

Hood, S. (2009). Evaluation for and by Navajos: A Narrative Case of the Irrelevance of Globalization. In *The SAGE international Handbook of Educational Evaluation,* edited by K. Ryan and B. Cousins, 447-464. Thousand Oaks, CA: Sage Publications.

Hood, S., Hopson, R., and Frierson, H. T. (eds.). (2005). *The Role of Culture and Cultural Context: A Mandate for Inclusion, the Discovery of Truth, and Understanding in Evaluative Theory and Practice.* Greenwich, CT: Information Age Publishing.

Jolly, E. J. (2002). On the Quest for Cultural Context in Evaluation: Non Ceteris Paribus. In *The cultural Context of Educational Evaluation: A Native American Perspective,* 14-22. Arlington, VA: National Science Foundation, Directorate for Education and Human Resources.

Kahle, J. B. (2000). Discussant remarks. In *The Cultural Context of Educational Evaluation: The Role of Minority Evaluation*

*Professionals* (NSF 01-43). Arlington, VA: National Science Foundation, Directorate for Education and Human Resources.

Kirkhart, K. E. (1995). Seeking Multicultural Validity: A Postcard From the Road. *Evaluation Practice,* 16 (1): 1-12.

LaFrance, J. (2004). Culturally Competent Evaluation in Indian Country. In *Search of Cultural Competence in Evaluation: Toward Principles and Practices*, edited by M. Thompson-Robinson, R. Hopson, and S. SenGupta, New Directions for Evaluation, No. 102, 39-50. San Francisco, CA: Jossey-Bass.

LaPoint, V., and Jackson, H. L. (2004). *Evaluating the Co-Construction of the Family, School, and Community Partnership Program in a Low-Income Urban High School.* New Directions for Evaluation, No. 101, 25-36. San Francisco, CA: Jossey-Bass.

Marin, G., and Marin, B. (1991). *Research With Hispanic Populations*. Newbury Park, CA: Sage Publications.

Mertens, D. M. (2003). The Inclusive View of Evaluation: Visions for the New Millennium. In *Evaluating Social Programs and Problems: Visions for the New Millennium*, edited by S. I. Donaldson and M. Scriven, 91-107. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Minkler, M. (2004). Ethical Challenges for the "Outside Research in Community-Based Participatory Research." *Health Education & Behavior*, 31(6), 684-497.

Nelson-Barber, S., LaFrance, J., Trumbull, E., and Aburto, S. (2005). Promoting Culturally Reliable and Valid Evaluation Practice. In *The Role of Culture and Cultural Context in Evaluation: A Mandate for Inclusion, the Discovery of Truth, and Understanding in Evaluative Theory and Practice*, edited by S. R. Hood, R. Hopson, and H. Frierson, 61-85. Greenwich, CT: InfoAge.

Patton, M. Q. (1991). Toward Utility in Reviews of Multivocal Literatures. *Review of Educational Research,* 61(3): 287-292.

Patton, M.Q. (2008) *Utilization-Focused Evaluation.* 4th Ed. Thousand Oaks, CA: Sage Publications.

Smith, L. T. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. London: Zed Books.

Stanfield, J. H. (1999). Slipping Through the Front Door! Relevant Social Science Evaluation in the People of Color Century. *American Journal of Evaluation*, 20(3): 415-431.

Stake, R. (1967). The countenance of Educational Evaluation. *Teachers College Record,* 68: 523-540.

Stake, R. (1980). Program Evaluation, Particularly Responsive Evaluation. In *Rethinking Educational Research,* edited by W. B. Dockrell and D. Hamilton. London: Hodder & Stoughton.

Stevens, F. I. (2000). Reflections and interviews: Information Collected about Training Minority Evaluators of Math and Science Projects. In *The Cultural Context of Educational Evaluation: The Role of Minority Evaluation Professionals* (NSF 01-43). Arlington, VA: National Science Foundation, Directorate for Education and Human Resources.

Thomas, V. G. (2009). Critical Race Theory: Ethics and Dimensions of Diversity in Research. In *The Handbook of Social Research Ethics*, edited by D. M. Mertens and P. E. Ginsberg, 54-68. Thousand Oaks, CA: Sage.

Thomas, V. G. (2004). Building a Contextually Responsive Evaluation Framework. In *Co-Constructing a Contextually Responsive Evaluation Framework: The Talent Development Model of School Reform,* edited by V. G. Thomas and F. I. Stevens, 3-24. New Directions for Evaluation, No. 101. San Francisco: Jossey Bass.

Thomas, V. G., and LaPoint, V. (2004/2005, Winter). Blending Evaluation Traditions: The Talent Development Model. *The Evaluation Exchange: A Periodical on Emerging Strategies in Evaluating Child and Family Services,* X (4): 7, 22.

Thomas, V. G., and McKie, B. K. (2006). Collecting and Utilizing Evaluation Research for Public Good and on Behalf of Africa American Children. *Journal of Negro Education,* 75(3): 341-352.

Thomas, V. G. and Stevens, F. I. (eds.). (2004). *Co-Constructing a Contextually Responsive Evaluation Framework: The Talent Development Model of School Reform*. New Directions for Evaluation, No. 101. San Francisco: Jossey-Bass.

Thompson-Robinson, M., Hopson, R., and SenGupta, S. (eds). (2004). *In Search of Cultural Competence in Evaluation: Toward Principles and Practices.* New Directions for Evaluation, No. 103. San Francisco: Jossey-Bass.

Worthen, B. R., Sanders, J. R., and Fitzpatrick. (1997). *Educational Evaluation.* 2nd Ed. White Plains, NY: Longman, Inc.

Zulli, R. A., and Frierson, H. T. (2004). *A Focus on Cultural Variables in the Evaluation of an Upward Bound Program.* New Directions for Evaluation, No. 102. San Francisco: Jossey-Bass.

# Chapter 8 ENSURING RIGOR IN MULTISITE EVALUATIONS

Debra J. Rog

## Introduction

Multisite evaluations (MSEs) are commonplace, especially in education (e.g., Lawrenz and Huffman, 2003). Although there is some literature on multisite methods (Turpin and Sinacore, 1991; Herrell and Straw, 2002), few discuss the strategies that can be used to ensure that the design, data collection, and methods are implemented with rigor.

This chapter provides guidelines for designing, implementing, and analyzing rigorous outcome MSEs as well as examples that illustrate the guidelines. It begins with a definition of multisite evaluation and advantages and disadvantages of conducting MSEs, followed by a review of different types of multisite approaches and dimensions on which designs can vary. The next sections go through key stages of an MSE, including developing the study foundation and initial design; designing the data collection methods and tools; assessing the interventions to ensure their integrity to expectations; devising strategies for maintaining the rigor of the study design and data collection, data management, synthesis, and analysis; and communicating the results. Many of the guidelines for MSEs are quite similar to those offered for any sound project evaluation, but the evaluation task is more complicated because of variation in contextual factors. Differences in local conditions may present challenges for implementing a one consistent evaluation design and this chapter suggests issues needing attention.

## Defining Multisite Evaluation

*Multisite evaluations examine a project or policy in two or more sites.*

Multisite evaluations examine a project or policy in two or more sites. In some instances, the intervention is intended to be exactly the same across the different sites. For example, classrooms implementing the mathematics program Knowing Math are expected to cover the same skills and complete the same number of units. In other situations, variations of an intervention are examined across multiple sites. Projects funded by the NSF's Math and Science Partnership Program (MSP), for example, typically include multiple sites, and these sites frequently vary in the activities in which they participate. What is

common among sites in the MSP is that each focuses on bringing about change in teacher knowledge, practices, and student achievement.[2]

**Advantages and Disadvantages of Multisite Evaluations**

MSEs have a number of advantages over single-site evaluations. Especially in new areas of intervention, they help to build the knowledge base more quickly than would otherwise occur. They often provide more powerful and influential results than single study evaluations, particularly when the data are collected with the same tools and measure and thus provide a "one voice" quality to the findings. Even descriptive findings can have more weight when they are collected, analyzed, and reported in the same manner across several sites. The research process also is likely easier to manage for an MSE than for a set of individual studies, especially if the project/research staff follow the same research principles, trial procedures, protocols, and guidelines. In addition, examining the implementation and outcomes of a project or policy across multiple sites that differ in geography, population composition, and other contextual features allows for increased learning about the generalizability of an intervention and its effects. Finally, there can be a value-added to an MSE that involves individual evaluations (experimental or quasi-experimental studies) across the different sites conducted by individual researchers who are also involved in the cross-site endeavor. When the collaboration among investigators is positive and active, it can become an "invisible college" (Reiss and Boruch, 1991) that builds the capacity of the evaluators across that sites and, in turn, improves the cross-site evaluation (Lawrenz, Huffman, and McGinnis, 2007).

*There can be value-added to an MSE that involves individual evaluations (experimental or quasi-experimental studies).*

MSEs also can have challenges. When there is less uniformity in the interventions across sites, the MSE can be left with analytic challenges due to these differences. Similarly, lack of standardization in designs and methods can complicate designs and analyses and make it difficult to draw conclusions. On the other hand, MSEs that strive for uniformity and standardization can have their own sets of challenges. It is more expensive to conduct highly collaborative MSEs than to evaluate individual sites separately, because they take more time to design, implement, and analyze due to the involvement of so many stakeholders and the need to consider the different points of view. Particularly when the design and methods are to be developed collaboratively, there can be philosophical and scientific disagreements that make it difficult to move ahead. Errors in an MSE, if not caught, can be more serious and long-lasting than errors in evaluating a single site because they are amplified by the number of sites (Kraemer, 2000).

[2] In this chapter, MSEs refer to the evaluation of a project that is implemented in multiple sites. This is different that an evaluation of a program that may includes multiple projects having similarities and differences.

## Multisite Approaches and Designs

There is no single outcome MSE approach or design. Rigorous MSE designs can vary on a range of dimensions, including the following:

- The nature of the individual site study designs—the sites can be all randomized studies, a mix of randomized and quasi-experimental studies, or all quasi-experimental studies;
- Treatment interventions—the individual studies can strive for identical interventions or include variations within a broader domain of intervention;
- The genesis of the interventions—they can be part of a demonstration or existing program initiative or constructed specifically for the MSE;
- Comparison interventions—the comparisons can vary by site or be constructed with identical procedures; and
- Sites selected for the MSE—either all the sites within a project or initiative or only a sample of sites can be included in the MSE. Samples can include representative sites, sites with a threshold level of fidelity, sites that are willing to be part of the MSE, and so forth.

### Factors That Determine the MSE Design

As noted above, the only distinction that MSEs share is that the evaluation examines an intervention in two or more sites. How the intervention is evaluated, however, depends on a number of factors, much like the design of a single-site evaluation.

Among the factors that shape the MSE design include:

- The nature of the evaluation questions(s);
- The nature of the problem that the intervention is addressing;
- The nature, diversity, and number of sites; and
- The resources (time, expertise, and funding) for the interventions/programs in each site, for the local evaluations (if applicable), and for the cross-site evaluation.

*A range of designs are often required to meet the variety of evaluation questions and purposes for the information.*

Therefore, as with single-site evaluations, a range of designs are often required to meet the variety of evaluation questions and purposes for the information. The nature, number, and diversity of the sites have a strong

bearing on the design that is desired and feasible. For example, if all study sites are expected to implement the same program, there likely will be attention to measuring the fidelity of implementation of the program. If, however, the programs are more diverse but fit under a more global program category, an examination of the sites may focus more on identifying the common features that the sites share as well as those on which they differ. Finally, the number of sites will likely influence the study design, especially in determining the nature of data collection and management. The larger the number of sites, the more important are standards for data collection, quality control, and data submission. With a large number of sites, the multisite evaluator may be in the position of determining whether all or a sample of the sites should be included in the evaluation.

### Sampling Sites

Many MSEs include all the sites in an initiative, but there are times when a sample is needed, either due to budget constraints or to the need to focus the evaluation either on the most rigorous sites or on some other selection criterion. For example, in projects that have a large number of sites, the MSE might include representative sites from clusters of sites sharing similar characteristics. In other MSEs, sites may need to meet a threshold level of fidelity of implementation of the intervention or demonstrate that they can successfully implement a randomized study or strong quasi-experimental study.

*In projects that have a large number of sites, the MSE might include representative sites from clusters of sites sharing similar characteristics.*

In some situations, there may be "sites within sites" that need to be sampled, such as schools within districts or classrooms within schools. For example, in an ongoing MSE of a science education reform project, the Merck Partnership for Systemic Change, activities to enhance the quality of science in instruction are being implemented in multiple schools in six different districts. Although the study is collecting data from teachers and principals in all participating schools, for cost concerns case studies of program implementation are restricted to a stratified sample of schools teaching certain courses and using pre-identified modules.

## Laying the Foundation for a Multisite Evaluation

*In MSEs that strive for rigor, there is a need to make certain that all study expectations, procedures, and developments are known by all key participants in the sites.*

Communication is the foundational element that is most distinct for an MSE compared to single-site evaluations. In MSEs that strive for rigor, there is a need to make certain that all study expectations, procedures, and developments are known by all key participants in the sites. In addition, for MSEs that are highly collaborative, it is important that participants in the local sites are clear on their roles in the cross-site study, participate in any training, and stay

connected to the cross-site evaluation. It is imperative that the MSE staff provide a variety of strategies for communicating to the staff in the local sites and obtaining their input and engagement. Communication mechanisms should include a mix of in-person, telephone, and electronic communications. Frequent communication and opportunities for sites to help shape some of the cross-site strategies decrease the potential for misunderstanding what is expected and, more importantly, build cooperation and good will, making it less likely that the local sites will go on their own or thwart the decisions that are made.

With improvements in technology, there is a growing range of mechanisms that can be used to both communicate and collaborate with local sites. Webinars, for example, can be used for training, and listserves and other email tools can be used for exchange of information. SharePoint, providing a "virtual filing cabinet," offers a central place for storing documents that can be accessed by anyone involved in the MSE. It has the added advantage of ensuring version control on key documents.

Interactive communication is often needed in MSEs that involve a high degree of collaboration and shared decision-making. All participants need to understand the specifics of all sites that compose the MSE and why certain cross-site design decisions are warranted.

## Multisite Data Collection: Developing a Common Protocol

When an MSE is a collaborative with individual site researchers working with the cross-site evaluation team, the first step will be to ensure that all sites agree with the research questions and overall framework. Building logic models (Frechtling, 2007) at this stage—ideally together with the site evaluators—and reviewing them with all key stakeholders will help to foster agreement on the main purposes and goals of the intervention, articulate the theory of change through the specification of short-term and longer term outcomes and, it is hoped, begin the process of delineating the measures needed to understand the implementation and outcomes of the project.

In addition to determining what needs to be measured, there is a range of decisions that must be made in developing a primary data collection protocol. Although most of these decisions need to be considered in single-site evaluations as well, in an MSE, the additional concern for each decision is whether and to what degree the decision has to be shared by all sites or can vary across the sites.

*In an MSE, the additional concern for each decision is whether and to what degree the decision has to be shared by all sites or can vary across the sites.*

Among the decisions that are considered in developing the protocol, researchers must first jointly determine the following:

- The population of interest and criteria for selecting participants;

- Strategies for recruiting and tracking participants, and the length of the recruitment period;

- Methods for collecting the data (i.e., in-person interviews, self-administered questionnaires, observational methods, web questionnaires);

- Logistics for data collection, i.e., whether data collection will involve the use of paper-and-pencil only or computers (e.g., CAPI, CATI, or CASTI), Optiscan, or web technology;

- Translation, i.e., whether and for which languages the instrument would need to be translated;

- The type of data collectors needed, i.e., whether outside data collectors will be needed and/or whether program personnel can collect any of the data; specific skills needed other than ability to collected standardized data (for example, or a need for bilingual or multilingual interviewers or a need for clinical training);

- Training for the data collectors, i.e., whether any specialized training is needed; and

- Timeframe for the data collection, i.e., whether there will be more than one wave of data collection and the timing of the waves.

Standardization on most decisions is usually preferred to ensure as much rigor as possible in the study procedures. The interest is in ensuring that the data collection strategies and the measures are the same or similar enough that if differences in the results do emerge across sites, they are not due to the methods being used. However, in some instances, having the same procedures across diverse sites can jeopardize feasibility and, at times, the validity of the data collection if all sites are not equally prepared to implement it. In such cases, a procedure may have to be adopted that is appropriate and can be used by all sites, even if it isn't the evaluator's first choice. That said, there are times when tailoring is the best solution. When translation is needed and there are multiple language groups involved, the process often requires tailoring the translation to specific dialects. One approach is to have a survey or instrument translated into generic Spanish, and then customized to the specific form of Spanish used in the particular sites (e.g., Puerto Rican; Mexican).

*Standardization on most decisions is usually preferred to ensure as much rigor as possible in the study procedures.*

Pretesting and piloting the cross-site tool and procedures is best done in all sites to determine how well it can work in each site and what modifications may be needed to tailor it appropriately to the site conditions.

## Assessing the Interventions

### Monitoring Fidelity

In MSEs in which a specific treatment intervention model is expected to be implemented in each of the sites, a fidelity assessment is performed to assess the extent to which this is true. A fidelity assessment typically involves the development of a tool that is guided by an understanding of the key elements and components of the program model. It can include looking for specific types of staffing, the level of implementation of different types of program components, and even the existence of different philosophical underpinnings within the program. Elements can be measured as to whether they exist or not, or rated according to the extent to which they are present.

Fidelity assessments can have multiple purposes. In some studies, measuring the extent to which sites have fidelity to a model can be included in the cross-site analyses. A second purpose for fidelity assessment is as a screening tool to determine whether sites have a sufficient level of the program to be included in the MSE. Finally, fidelity assessments can be incorporated within formative stages of an evaluation to inform mid-course corrections in program implementation.

### Assessing Comparison as Well as Treatment Sites

When an MSE involves comparison sites that have a comparison intervention, it is useful to apply the same or similar emphasis on understanding the nature and strength of the comparison interventions as applied to the treatment interventions. In such cases, it may be important to determine if the comparison sites have fidelity to the treatment alternative. Examining the comparison conditions thus can provide an understanding of the extent to which they provide an adequate contrast with the treatment conditions as well as whether what is being compared is consistent with design intentions. If data on the comparison conditions are obtained early enough, this knowledge can help to refine the design. Obtaining the information at any time can help to shape analyses as well as assist in properly interpreting the final results.

*It may be important to determine if the comparison sites have fidelity to the treatment alternative.*

## Maintaining the Rigor of the Study Design

Part of the role of the MSE is ensuring that the individual study designs are being implemented with integrity, especially when the same design is expected in all sites. When the site-level investigators are in control of the design decisions, it is important at the MSE level to understand the decisions that are being made and the nature of the design that is being implemented. For example, sites may differ in how they construct and

implement comparison groups, and this variation will affect the analysis and possibly the results.

MSEs also need to monitor the implementation of agreed-upon procedures related to participant selection, recruitment and tracking, as well as data collection and the logistics involved with data collection (described more completely in the section on quality control). At times, contextual differences in the sites may influence the degree to which these procedures can be implemented and the extent to which modifications are needed. In other sites, shifts in funding may create changes in the program and may also affect the participants. Although these changes may be inevitable, the key to maximizing rigor and integrity in an MSE is trying to have the response to these changes be as uniform as possible across sites.

Multiple monitoring strategies are often needed, especially with complex studies and interventions. In addition to frequent contact with the individuals in the local site responsible for design and data collection, other monitoring methods include regular site visits to review study procedures, status reports on study implementation developed and submitted by the sites, group conference calls, email exchanges, and in-person meetings that offer opportunities to share information on implementation problems and develop shared solutions to the problems.

## Quality Control in MSE Data Collection

*Ensuring rigor in an MSE necessitates having quality in measurement.*

Ensuring rigor in an MSE necessitates having quality in measurement. Quality control procedures are needed to ensure that there is uniformity in data collection. These procedures fall into two main areas: selecting, hiring, and training data collectors; and ongoing review of data collection.

### Selecting, Hiring, and Training Data Collectors

Selecting and hiring data collectors for MSEs lies in the domain of the MSE evaluation team, and it is important that there be a discussion of the criteria that should guide selection and hiring. For some MSEs, it may be important that the data collectors have certain credentials and prior training (e.g., ability to know what is accurate mathematics content in a classroom observation) as well as certain characteristics (e.g., preference for ethnic similarities in the data collection staff and the study population). Given that sites may vary in the extent to which a local pool of data collectors is available, it is important that the selection criteria be those that are considered essential for the effort, not just perceived to be desirable. Over-specifying the criteria can make it difficult for some sites to find individuals that meet them. In addition, there may be some site-specific considerations. Data collection training is best conducted

centrally in an MSE to ensure that all are hearing the same story. If there are a large number of data collectors across the sites, a train-the-trainers session can be conducted. For this type of session, each site designates one or more individuals to attend the cross-site training who then serves as the trainer for the rest of the data collectors in the site. To maintain the cross-site integrity in the data collection and associated procedures in studies that will be conducted over a period of time, if the budget can support it, it is best to have multiple people from each site attend the training in the event of turnover. Additional strategies for assuring uniformity in training include videotaping the central training and having it accompany a live training in each site, and holding Webinar trainings in which all site trainers and/or all interviewers participating in the central training can participate.

The nature of the training is generally similar to what would be covered in a single-site evaluation—the basics of interviewing; how to respond to unusual interview situations; obtaining informed consent; and the specifics of the data collection instrument, including the nature of the domains and measures, how to follow skip patterns, how to select individuals on whom to collect data (if relevant), and how to collect information to help in tracking the respondent for future data collection. In addition, it is also customary (especially in interviews) to include a section that the data collector completes indicating his or her assessment of the validity of the data, given other contextual factors (e.g., distractions, visible concerns of the respondent for confidentiality, etc.).

In longitudinal MSEs, booster training sessions on data collection can reinforce aspects of the original training. Boosters can help interviewers and other types of data collectors avoid ruts as well as cover upcoming follow-up tools or changes that are occurring in procedures due to site changes or other unforeseen issues.

In some data collection efforts, sections of the interview or process are less standardized but need to follow certain procedures. In these instances, a readiness assessment can be conducted to determine if a data collector can follow these sections of the protocol before going out into the field. A "gold standard data collector," typically the trainer at a site, is the individual who has mastered the data collection process and thus is the one against whom all other data collectors are assessed.

For data collection efforts that span a period of time, routinely scheduled reliability assessments are recommended. For interviewers, for example, these assessments involve randomly selecting interviews to be audiotaped and subsequently reviewed to determine if the data collection and coding procedures were followed. MSE staff typically conduct the reviews and communicate the results to the data collector and the supervisor in writing, highlighting the areas that the data collector followed as expected as well as areas where there were slippages and retraining or refreshing may be needed. The process also identifies data collectors who

*For data collection efforts that span a period of time, routinely scheduled reliability assessments are recommended.*

consistently have trouble with the procedures and may need more intensive training or dismissal.

### Ongoing Data Collection Review

A key role of the MSE team is ensuring the quality of the data collected across the sites. In addition to the quality assurance activities incorporated into training, ongoing data collection review demands a number of strategies, from regular communication with staff in each site to central review and analysis of the data on an ongoing basis. Site data coordinators, either hired by the MSE directly or on the staff of the local site, can serve as the key contacts with the MSE team. The data coordinators, who may also be gold standard data collectors are responsible for monitoring the data collection effort and serving as the supervisors. They need to be up to date on all procedures so that they can ensure the data collectors are collecting the data as intended. They also are the main source of information on the status of the data collection effort, including observations completed, surveys returned, the recruitment and interview rates, problems in obtaining context, and so forth. In addition, in fielding concerns from the data collectors especially in the initial stages of the data collection, they will identify situations that are not covered by the training or data collection procedures. Rather than resolve them on their own, site coordinators need to be instructed to bring the problems to the attention of the MSE team so that any resolution can be uniform across the sites. These resolutions can then translate into "decision rules" that can be added to the training materials as well as maintained on SharePoint or some other vehicle that allows for easy access across the sites.

Once data are submitted to the cross-site team, the data should be reviewed quickly by the cross-site team to confirm that they are being collected and coded as expected. Early reviews can identify areas that are not being followed according to directions or coding decisions that were not fully explicated.

## MSE Quantitative Analysis

### Preparatory Steps

As with any evaluation, several important preparatory steps are needed in MSE data analysis. They include data cleaning and manual review when the data are collected on hard copy, and computerized data cleaning for all data submitted to assess validity and accuracy. Also important are analyses that examine the quality of the data such as testing for floor and ceiling effects, patterns of nonresponse or responses that demonstrate lack of understanding of the questions, and consistency checks among the items. In addition, in longitudinal studies, it is sometimes necessary

to assess and control for artifacts (such as attrition) or timing differences in the completion of follow-up interviews.

Cross-site tables and graphical analysis (Henry, 1995; Tufte, 2001) can be especially useful in MSEs for examining early data problems and patterns in the data as well as elucidating differences and similarities across the sites. In the initial analysis stage, graphical displays, such as scatterplots that display the variation within and across the sites, can be very useful. Box-and-whisker diagrams, for example, can readily show differences among sites in the distributions to an item or scale, in the display of the minimum score, the lowest quartile, the median, the upper quartile, and the maximum score, as well as any outliers. Star plots also can be used to show differences in frequency distributions for multiple sites on multiple variables.

**Pooling Data**

A significant advantage of MSEs, when it is appropriate, is to pool the data across the project sites and, in turn, to have greater statistical power than would be achieved with any single site. In addition, with pooled data, MSEs can take advantage of a variety of multivariate techniques, some of which are specifically designed for nested data. Multilevel models in particular allow for separation of the variance due the site from the variance due to the individual participants. These models thus provide the ability to examine the effects of an intervention on its participants with proper analytic controls on the appropriate site variables (e.g., schools, classrooms, students).

Having larger samples across sites provides the ability to examine outcomes for key subgroups of study participants that would otherwise be too small in any one site, for example, for individuals from different racial and ethnic categories. It also provides the ability to look at the role of context in shaping outcomes, for example, examining the differences in outcomes for children from rural vs. urban sites, as well as from states or communities with different levels of social capital.

For longitudinal studies, one of the strengths of a pooled data set is the ability to see if there are different patterns of change among the participants and how those patterns relate to the interventions as well as individual characteristics and contextual features. These trajectories can be analyzed using trajectory analysis (Nagin, 1999) or growth-mixture models (e.g., Muthen and Muthen, 2000) and are powerful in getting closer to answering the key question "what works for whom under what circumstances?"

*The more diverse the site . . . the more challenges there are likely to be in the analysis.*

However, there may be analytic challenges posed by the sites that make pooling difficult or impossible. Clusters of sites may also be pooled if there are greater similarities in subsets of the sites than across all sites. The more diverse the sites are (i.e., in the populations served, the measurement and data collection

processes, the programs evaluated, and the evaluation contexts) and the less cross-site control there is, the more challenges there are likely to be in the analysis. It may be useful, therefore, to build contingencies, such as conducting individual site analyses and meta-analyses (Banks, et al., 2002) of the data into the plan.

### Maintaining Independence in the Data

In situations in which the treatment conditions being tested or the populations in each study appear too distinct, prospective meta-analysis (Banks et al., 2002) may be an alternative to pooling. Prospective meta-analysis involves meta-analytic methods that are usually used to statistically integrate the findings from a set of existing studies all addressing the same general research question but typically with different populations, situations, etc. As with standard meta-analysis, the metric used in prospective meta-analysis is the effect size, defined as the standardized difference in outcome between the average intervention study participant and the average comparison group study participant.

### Design Sensitivity

In some MSEs, it is useful to use a design sensitivity approach to the data analysis. Design sensitivity, coined by Mark Lipsey (1990; Lipsey and Hurley, 2009) refers to maximizing a study's statistical power. Lipsey's approach to study design is to focus on those factors that influence statistical power, such as the strength and integrity of the treatment intervention; the level of contrast between the treatment and control conditions; the size and homogeneity of the study sample; the quality of the measurement and its sensitivity to change; and statistical analyses that optimize statistical power, such as the use of blocking variables.

*Design sensitivity refers to maximizing a study's statistical power.*

At the analysis stage of an MSE, a sensitivity approach considers these factors and how they might be "strengthened" in the analysis to see if the intervention may be having an effect that is obscured by variability in the design. For example, in large multisite studies with some variation among the sites, sensitivity analyses might include only those sites that have high fidelity to the treatment and/or participants who have received a threshold level of an intervention, or only sites in which there is a sufficient contrast between the treatment and control conditions.

## Qualitative Analysis Strategies

Qualitative analysis strategies for MSEs are similar to those that would be used in any individual study. However, given the large scope of most MSEs and often limited timeframes, exploratory or grounded approaches

to qualitative analysis are usually difficult to do well. More upfront structure is often needed in the MSE qualitative data collection and analysis to ensure that the potentially large volume of data can be collected and analyzed in a timely and cost-efficient manner and within budget.

Analysis involves a successive series of data reduction steps. For example, understanding the level of implementation of each site is likely to entail collecting a range of qualitative data through site visit interviews, document reviews, and observations according to a set of domains detailed in a data collection protocol. Data reduction is likely to begin with summarizing the data on each implementation domain by the source (e.g., project director interview), then across sources (e.g., all interviews, documents, and observations) and possibly all implementation domains to reach an overall assessment of a site's implementation. After performing this set of steps for each site, the evaluator then needs to compare and contrast the implementation level of all sites.

Software packages, such as NVivo and Atlas can be used to help organize qualitative data. These programs can also be used to perform the analyses and integrate data from multiple media.

Data displays that array data by different dimensions (such as by chronological time) can also be useful in performing qualitative analyses by helping to illustrate patterns in the data within sites and across sites (Miles and Huberman, 1994).

**Strategies for Reporting and Briefing**

Maintaining rigor in MSE reporting means ensuring that the cross-site messages of the data are accurately communicated, while not obscuring the key site differences. Graphs and tables are critical to presenting complex findings, especially where there are differences among sites or key patterns. In past studies, we have been successful in developing "dot" charts that display the presence or absence of a dimension in a site and can be scanned very quickly to see cross-site patterns in the data (see Rog et al., 2004 for an example of a dot chart).

*Maintaining rigor in MSE reporting means ensuring that the cross-site messages of the data are accurately communicated, while not obscuring the key site differences.*

In highly collaborative MSEs, it is important to have individuals from across the sites involved in the interpretation of findings and in the crafting and/or editing of the report. Joint authorship, however, requires early policies that state who is permitted to issue press releases, prepare publications, and otherwise report the findings of the MSE and when those communications can be made.

## Conclusion

MSEs are increasingly common in education and a variety of areas. There is no one type of MSE but, rather a range of designs and approaches that can be used. This chapter has attempted to provide a portfolio of useful approaches and designs and strategies for ensuring rigor in whatever type of MSE is employed. The strategies span the study process from design and data collection through data management, to synthesis and analysis, and finally to communication of the results.

## References


Banks, S., McHugo, G. J., Williams, V., Drake, R. E., & Shinn, M. (2002). A Prospective Meta-Analytic Approach in a Multisite Study of Homelessness Prevention. In *Conducting Multiple Site Evaluations in Real-World Settings,* edited by J. M. Herrell and R. B. Straw. New Directions for Evaluation, No. 94, San Francisco: Jossey-Bass.

Frechtling, J. (2007). *Logic Modeling Methods in Program Evaluation.* San Francisco: John Wiley and Sons.

Henry, G. T. (1995). *Graphing Data: Techniques for Display and Analysis*. Thousand Oaks, CA: Sage.

Herrell, J. M., & Straw, R. B. (2002). *Conducting Multiple Site Evaluations in Real-World Settings.* New Directions for Evaluation, No. 94, San Francisco, CA: Jossey-Bass.

Kalaian, S. A. (2003). Meta-Analysis Methods for Synthesizing Treatment Effects in Multisite Studies: Hierarchical Linear Modeling (HLM) Perspective. *Practical Assessment, Research, and Evaluation*, 8 (15).

Kraemer, H. C. (2000). Pitfalls of Multisite Randomized Clinical Trials of Efficacy and Effectiveness. *Schizophrenia Bulletin, 26*, 533-541.

Lawrenz, F., and Huffman, D. (2003). How Can Multi-Site Evaluations be Participatory? *American Journal of Education*, 24 (4), 471-482.

Lawrenz, F., Huffman, D., & McGinnis, J.R. (2007). Multilevel Evaluation Process Use in Large-Scale Multisite Evaluation. In *Process Use in Theory, Research, and Practice*, edited by J. B. Cousins, 75-85. New Directions for Evaluation, No. 116, San Francisco: Jossey-Bass.

Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage.

Lipsey, M. W., & Hurley, S. M. (2009). Design Sensitivity: Statistical Power for Applied Experimental Research. In *The Sage Handbook of Applied Social Research Methods*, edited by L. Bickman and D. J. Rog. Thousand Oaks, CA: Sage.

Miles, M., and Huberman, M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, CA: Sage.

Muthen, B., & Muthen, L. (2000). Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling With Latent Trajectory Classes. *Alcoholism: Clinical And Experimental Research*, 24, 882-891.

Nagin, D. S. (1999.) Analyzing Developmental Trajectories: A Semi-Parametric, Group-Based Approach. *Psychological Methods*, 4, 139-177.

Reiss, A. J., & Boruch, R. (1991). The Program Review Team Approach and Multisite Experiments: The Spouse Assault Replication Program. In *Multisite Evaluations*, edited by R. S. Turpin & J. M. Sinacore. New Directions for Program Evaluation, No. 50. San Francisco: Jossey-Bass.

Rog, D., Boback, N., Barton-Villagrana, H., Marrone-Bennett, P., Cardwell, J., Hawdon, J., Diaz, J., Jenkins, P., Kridler, J., and Reischl, T. (2004). Sustaining Collaboratives: A Cross-Site Analysis of The National Funding Collaborative on Violence Prevention. *Evaluation and Program Planning*, 27, 249-261.

Tufte, E. (2001). *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

Turpin, R. S., & Sinacore, J. M. (Eds.) (1991). *Multisite Evaluations.* New Directions for Program Evaluation, No. 50, San Francisco: Jossey-Bass.

Chapter 9

# PROJECT EVALUATION FOR NSF-SUPPORTED PROJECTS IN HIGHER EDUCATION

Melvin M. Mark

The good news is that you've been asked to write the evaluation section of a proposal to NSF for a project designed to enhance STEM higher education. Of course, you don't want your evaluation plan, and the response grant proposal reviewers will have to it, to be the bad news. So how do you proceed?

Or the good news may be that you plan to submit a proposal to NSF, building off of work you've been doing in STEM education. Even better, you may have received encouragement about your planned proposal from an NSF program officer. The bad news, however, is that you are required to include an evaluation plan, and evaluation is not your specialty. You've talked with someone in the local college of education who you might bring on as the project evaluator, but you would like to be a bit more comfortable yourself with what the options and best course of action are. So how do you proceed?

It might seem as though the answer should be easy. If you are the evaluator-to-be, you probably are coming to the evaluation with expertise in a certain kind of research. You may feel that your task is to figure out how best to link your skills to the project evaluation. For example, perhaps your background is in measurement. If so, you may expect that your challenge in designing the project evaluation will be to identify what concepts need to be measured and then to create and implement a plan for developing high-quality measures for use in data collection. Or perhaps you are a proficient leader of focus groups. If so, your initial thought may be that your evaluation planning will involve figuring out whom you need to get to participate in focus groups and on what topics.

Alternatively, perhaps as the project PI you think that planning the project evaluation should be straightforward for a very different reason. You may have heard that one kind of research design has been called the "gold standard" for evaluation. You may be vaguely familiar with arguments for more rigorous evaluation designs, where the concept of rigor is equated with the use of those "gold standard methods" or their closest cousins. (We'll return later to the details of the gold standard discussions.) Or, as the prospective evaluator, you may also be aware of organizations that summarize evaluation findings with methodological "screens," that is, rules that keep evaluations with certain kinds of designs in the summary while excluding others. Having heard about such things, you may assume that the project evaluation should use methods that are as close as possible to the kind some people consider to be gold standard.

*One size does not fit all. That's true whether the "one size" is based on your own expertise or on some generic claim about gold standards.*

The only problem with any of these approaches is that they are likely to lead to the wrong kind of project evaluation in many, if not most, cases. One size does not fit all. That's true whether the "one size*" is based on your own expertise or on some generic claim about gold standards. Admittedly, it's easy to say that one size does not fit all. But how should you go about trying to ensure that the planned evaluation fits well with the project being evaluated? Put differently, how can you think smarter about evaluating an NSF-funded project that is aimed at improving STEM education? This chapter discusses factors to consider in designing your STEM evaluation plan. While the focus is on evaluation programs in higher education, the principles discussed apply broadly across the PK–20 education spectrum.

## An Early Consideration: Evaluation Purpose

*An early and important task . . .is to figure out which of the potential benefits of evaluation are most important for your project.*

One of the first things to think about is *why* the project will be evaluated. Put differently, what purpose or purposes is the evaluation intended to meet? Of course, a pragmatic answer can be given: NSF requires an evaluation, and the project won't be funded without one. However, this requirement exists for a reason. NSF believes that in general, project evaluation will result in one or more kinds of benefits. An early and important task in evaluation planning is to figure out which of the potential benefits of evaluation are most important for your project.

As described elsewhere in this Handbook, evaluation can have different primary purposes. One common purpose corresponds to what is called "formative evaluation." A formative evaluation is intended to help improve the thing being evaluated. Take as an example a formative evaluation of a new project funded under NSF's Research Experiences for Undergraduates (REU) program. The evaluation's purpose would be to help project staff improve the local REU project. This could include such things as: identifying better ways of recruiting eligible students; discovering obstacles that keep potential research mentors from participating; and examining the apparent strengths and weaknesses of the way this REU project is implemented. To carry out a formative evaluation like this, the evaluator might start by working with project staff to develop a logic model (described in Section II). Such an effort could reveal any apparent gaps in the project's rationale and lead to the creation of a better project plan. The evaluator might also observe the project in operation. Finally, the evaluator might interview individuals from several groups, including project staff, research mentors, undergraduate student participants, and some potential mentors and eligible students who did not participate.

A second common purpose of evaluation corresponds to what is called "summative evaluation." A summative evaluation is intended to provide a bottom-line judgment about the thing being evaluated. For instance, a follow-up project might have been funded under NSF's Course, Curriculum, and Laboratory Improvement (CCLI) Program.[3] Among other things, funded projects that assess the effectiveness of educational tools. Imagine, for instance, an online tool called FrankenGene that allows students to simulate transgenetics projects they think would be interesting. In transgenics, a gene from one kind of organism is transplanted into another. To take an actual example, a gene that makes jellyfish fluorescent has been transplanted into a pig, making a pig that glows in the dark. Imagine further that the online FrankenGene tool had been developed under a previous NSF grant that had emphasized formative evaluation. In this case, in the newly funded project the evaluator would focus on a summative evaluation. Perhaps this would take the form of an experiment designed to see what effect, if any, use of the website has on students' learning, their interest in STEM professions, and other outcomes of interest. In essence, students who use FrankenGene would be compared with similar students who did not have access to the online tool in terms of their learning outcomes and their interest in STEM majors and careers.

*The purpose of an evaluation should derive in part from the project, what it is intended to achieve, and the questions it is addressing.*

From these two simple examples, one involving formative evaluation and the other summative evaluation, three key points emerge. First, the purpose of an evaluation is one of the considerations that should guide evaluation design. Second, in general, the kind of evaluation methods that make sense will likely be different for one evaluation purpose than for another. The experiment that makes sense for the FrankenGene summative evaluation might be of little if any value for the REU formative evaluation. Third, the purpose of an evaluation should derive in part from the project, what it is intended to achieve, and the questions it is addressing. For example, consider a project designed to assess the effectiveness of a previously developed educational product before it is widely disseminated. In such a case, it makes sense for the evaluation to be summative, aimed at giving a thumbs up or thumbs down judgment. Why? Because summative evaluation findings could inform relevant future action, specifically by clarifying whether widespread dissemination would be a good idea. On the other hand, for a new REU project, the first order of business likely would be more formative, aimed primarily at program improvement rather than at a confident bottom-line judgment.

Three additional points may not be evident from the preceding examples but will be elaborated upon in the remainder of this chapter. First, choice of evaluation purpose is typically a matter of degree, and sometimes one of timing, rather than an all-or-nothing proposition. For example, when

[3] Recently NSF changed this program to Transforming Undergraduate Education in Science, Technology, Engineering and Mathematics (TUES).

evaluating a new REU project, the emphasis may primarily be on formative work. Still, some data related to a tentative summative judgment, such as participating student's interest in STEM careers, would also be collected. Across time, a product such as FrankenGene typically would be improved with formative evaluation prior to an intense (and perhaps relatively expensive) summative evaluation. Second, a combination of methods will usually be preferable to a single method in project evaluation. For example, when conducting an experiment to test a new educational tool such as FrankenGene, one might also conduct classroom observations to see how the tool was actually used in practice. Third, evaluation design typically involves pragmatic considerations and tradeoffs. To take a simple but potent example, the size of the evaluation budget affects how much can be done.

## Evaluation Design: It Depends

*"What does the 'best' or 'most rigorous' evaluation look like…?" The right answer, as the previous discussion suggests, is, "It depends."*

Consider the question "What does the 'best' or 'most rigorous' evaluation look like for a STEM education project supported by NSF?" The right answer, as the previous discussion suggests, is "It depends." However, "it depends" is not by itself a terribly satisfying answer. More specific guidance, or at least a way of thinking about the options, is needed. If the remainder of this chapter meets its goal, it will help you do better at working through the process of considering and choosing from among several options for project evaluation.

We've already considered the idea that evaluation purpose is one of the factors that should affect what an evaluation looks like. Evaluation has potential purposes beyond the classic distinction between formative and summative. Indeed, it seems that every other book on evaluation today contains an argument for a new evaluation purpose that evaluation might strive to achieve. Fortunately, not every potential evaluation purpose is likely to be central to NSF projects evaluations. For example, some private foundations have tried to use evaluation as a vehicle for improving overall management capacity in the nonprofit organizations that they fund. This kind of overall organizational capacity building probably will not be a central purpose for most NSF-funded projects that support institutions of higher education.

However, other potential evaluation purposes may be relevant to NSF project evaluation.

- One potential purpose of evaluation is to meet *accountability requirements*. For example, NSF's Louis Stokes Alliances for Minority Participation (LSAMP) Program is designed to foster involvement in STEM disciplines by members of traditionally

underrepresented groups. An evaluation of an LSAMP project could satisfy an accountability purpose, for example, by tracking the number of students at the project institution who are involved in project activities, as well as the number from minority groups who are majoring in STEM disciplines over time. Sometimes accountability demands are as simple as knowing how many people received project services, and that they met any eligibility requirements.

- Another potential purpose of evaluation involves *contributing to the knowledge base*. Knowledge development often is not a sole evaluation purpose. Rather, new knowledge may be sought in conjunction with another purpose. The LSAMP proposal solicitation, for instance, states that proposals will be judged in part based on the likelihood that the project evaluation will contribute "to the body of knowledge in transforming student learning, recruitment and retention of underrepresented minorities in science, technology, engineering, and mathematics disciplines and into the workforce" (NSF 08-545, p. 12).

- Summative evaluation of NSF projects typically relies on examining project-related outcomes. For example, do students who use a web-based tool perform better on a test of learning than students who do not have access to the tool? However, evaluation can also focus on the *feasibility of implementing a new approach*. With a new teaching technology, or a radically different approach to attracting minority students' participation in STEM, sometimes the key first question is "Can you actually do it this way?" In other words, evaluation can serve the purpose of offering basic *proof of concept* or clarifying ways in which a concept might be flawed. Another purpose advocated by some evaluation scholars and practitioners is to understand *participants' lived experience*. This may be less likely than other evaluation purposes to make sense for an NSF project evaluation, but in certain cases it may be appropriate. Imagine, for example, a project at a predominantly white university that is designed to increase the pipeline of minority students into STEM professions. In this example, it could prove quite useful to understand what the project and its activities feel like to the intended beneficiaries.

*The nature of the evaluation should depend on the evaluation purpose or purposes that have priority for a given project.*

So, several evaluation purposes exist. Moreover, the nature of the evaluation should depend on the evaluation purpose or purposes that have priority for a given project. But how can one best choose from among the various evaluation purposes?

The purpose or purposes of evaluation that will predominate in a given project evaluation will depend on several factors. Usually one of these is the *stage or maturity* of the thing being evaluated. For example, if FrankenGene is not yet operational, then the evaluation purpose will be more formative and perhaps proof of concept. In contrast, if the website is fully functional and its advocates are interested in disseminating it

widely, then a more summative purpose, perhaps combined with accountability, would seem to be in order. The expected evaluation consumers' needs and potential uses constitute another factor that often influences the choice of evaluation purpose. For example, NSF staff members routinely need to report to Congress on NSF's programs, leading them to need information they can use to meet basic accountability purposes. (For many projects, this need will be met indirectly, through the project's contribution of information to the overall program evaluation.) In the early stages of an REU project, project staff are interested in ways to improve the REU, so they need the data-driven feedback that an evaluator can provide about the strength and weaknesses of the project. The general *state of knowledge* about the thing being evaluated can also influence the selection of evaluation purpose. For example, imagine that we are late in a series of NSF-funded projects involving FrankenGene. Also imagine that previous evaluations have already demonstrated that use of the website causes improvements in the outcomes of interest. With this as background knowledge, any subsequent evaluation might focus more on knowledge development as a purpose, say by studying *why* this web-based tool is effective.

In planning an evaluation of an NSF-funded project, the evaluator and project investigators are not in the position of trying to select evaluation purposes in a vacuum. Rather, one of the key influences on evaluation design should be NSF itself, particularly the program solicitation under which your proposal is being submitted. A good program solicitation will incorporate current thinking about evaluation at NSF, drawing on the experiences NSF staff have had with many related project evaluations. The program solicitations draw on NSF's thinking about the relative priorities for different kinds of evaluation purposes, presumably based on factors such as project stage, information needs and potential uses, and the state of relevant knowledge.

*One of the key influences on evaluation design should be NSF itself, particularly the program solicitation.*

Take as an example the solicitation for NSF's CCLI program (NSF 09-529). The goal of the program is to improve the quality of STEM education for undergraduates. Of particular interest are proposals that address learning materials and teaching strategies that have the potential to transform STEM education for undergraduates. The solicitation specifies that all project evaluations should include both formative evaluation and summative evaluation. The solicitation also describes three types, levels, or stages of projects. These three vary in terms of (a) the number of schools, faculty, and students involved, (b) the number of components being investigated, and (c) the maturity of the approach being studied. In general, the proportion of effort that is to be given to summative evaluation, relative to formative evaluation, increases as the project becomes larger, the intervention more multifaceted, and the approach more mature.

The solicitation also describes various foci, or components, that a proposal may have, and it specifies what the project evaluation should address for each component. For example, proposals may focus on how best to implement new instructional strategies. According to the solicitation, "Evaluation plans for implementation projects should explore the challenges and opportunities for adapting new strategies in diverse educational settings" (p. 4). In a sense, the basic implementation project under CCLI includes proof of concept as an evaluation purpose for innovative implementation strategies. The solicitation continues: "Projects that specifically address the challenges for achieving widespread adoption of proven practice are especially welcome" (p. 4). In essence, the solicitation asks for summative evaluation of well-formed approaches for the widespread dissemination of effective practices.

With another possible component, CCLI projects can focus on developing faculty expertise. According to the solicitation, such projects "should include evaluation efforts to describe the impact on the faculty participants, and in large, later stage projects on student learning in classes taught by these faculty" (p. 4). Note that the program solicitation calls for the summative component of the evaluation to vary depending on the project stage. For a project with a newer intervention, a less intense summative evaluation will focus on a shorter-term outcome, that is, whether faculty learn relevant knowledge and skills as a result of the project. A test administered to faculty before and after participation in the project's activities may well suffice. In contrast, for a more mature intervention, the summative evaluation needs to shift to a longer-term outcome, that is, student learning. In this case, the evaluation would probably compare student performance before and after a year in the classrooms of teachers who participated in the project activities, relative to the gains in classrooms of similar teachers who did not participate in the project. This might involve the use of an experimental design, described later.

Yet another component includes the development of new learning materials and strategies. With this focus, "Early stage projects typically carry the development of materials, and assessment of learning, to the stage where judgments can be made about whether further investment in the new materials or approaches is justified. Later stage projects should yield evaluation results sufficiently conclusive and descriptive so that successful products and processes can be adopted, distributed widely or, when appropriate, commercialized" (p. 4). Thus, the solicitation suggests a different standard for the summative evaluation depending on the project's stage. Earlier projects are held to a lesser standard of conclusiveness, relative to later stage projects. Also note that descriptive information is to be included in the summative evaluation of a later stage project. Presumably this will include information on how the learning materials are used in practice, so that later adopters can model the approaches used when the product was found to be effective. As this example suggests, multiple methods will often be required to maximize the value of an evaluation.

## Thinking about Tradeoffs

Part of the art of evaluation practice involves thinking through the tradeoffs that almost inevitably arise in practice. It may be enticing to think your project evaluation will be the one that does it all. The risk, however, is that by trying to do too many things, you end up doing nothing well. In principle, a comprehensive evaluation is possible. However, it would take considerable resources, including time. For most individual project evaluations, choices must be made and priorities established. Comprehensiveness should be an aspiration for a set of evaluations over time, if the work evolves into a series of projects that develop in maturity and scope.

*Part of the art of evaluation practice involves thinking through the tradeoffs that almost inevitably arise in practice.*

Huey Chen (2004) has given us one general way of describing the tradeoffs faced in evaluation. He says that four factors—breadth, rigor, cost, and time—commonly are in conflict with one another. Ask an evaluator to increase breadth, say, to conduct both summative and formative evaluation while also attending to proof of concept and knowledge development. If you don't provide the evaluation with considerable time and money, rigor almost certainly will suffer. For example, the summative evaluation would not be done as well as if it had been the only evaluation purpose. In contrast, if you increase the budget and extend the time frame of the evaluation, it should be easier for the evaluator to add breadth while maintaining rigor. If time and cost are both rather limited, as will be the case for many project evaluations, the tradeoff between breadth and rigor may be central to evaluation planning.

How best can you deal with tradeoffs? One way is to take into account another factor, specifically, what degree of confidence is needed for the evaluation findings. How good an answer is required—a general ballpark answer or a very precise answer about which you are quite confident? This may differ for the various evaluation purposes to be addressed in an evaluation. For example, for a primarily formative evaluation, the formative component may need to be quite strong, while a ballpark answer may suffice for the summative component. The converse may be true for a primarily summative evaluation. Similarly, the earlier in the development cycle an educational product is, the less confidence will generally be needed in summative evaluation findings. This kind of thinking was built into the solicitation, most explicitly for projects developing new learning material and strategies.

*Certain evaluation activities can help meet multiple purposes, if used judiciously.*

Another way to deal with tradeoffs is to try to avoid them. Certain evaluation activities can help meet multiple purposes, if used judiciously. For example, for a project with multiple components, the evaluator might develop a set of databases that track students' participation in various project activities as well as student outcomes such as degree completion and initial employment. The databases can be used for formative evaluation, such as by seeing whether certain project

components are not attended well. The databases can serve accountability as an evaluation purpose, enabling the project to report out to NSF program officers when information is requested. And the databases offer a degree of summative evaluation, providing data on project participants' early educational and employment outcomes.

## A Brief Review of the "Gold Standard" Debate

In the last decade or so, the notion of gold standards has been bantered about in discussions about how to do evaluation. In essence, the argument has been that randomized experiments, often labeled RCTs for "randomized controlled trials," provide the most rigorous method for evaluation. The argument further is that if a randomized experiment is not feasible for either ethical or pragmatic reasons, then one of its closest cousins should be implemented. Because the gold standard debate may be brought up in discussions about NSF project evaluations, it is useful to have a sense of what the debate has been about. A centrist perspective is presented here, and parties on either side of the debate may hold views that differ.

First, a bit of history. Randomized experiments have long been common in many areas of applied social research, including evaluations, such as studies of the effects of psychological treatments for anxiety and depression. In contrast, in the last decade, critics have argued that randomized experiments have been woefully underused in other areas, including education and international development. Moreover, this position has been translated into action in some grant-making and literature-reviewing processes. Certain funding streams at the Institute for Education Sciences at the U.S. Department of Education, for example, give priority to proposals with a randomized experiment. The What Works Clearinghouse (WWC) was designed to offer educators and others a summary of the best evidence about the effectiveness of educational products, programs, and policies. The WWC uses methodological screens such that RCTs and their closest cousins are included in its reviews, while evaluations using different methods are excluded. Advocates have often used terms such as "gold standard," "rigor," and "scientific" to describe randomized experiments.

The problem with such language is that without careful caveats, it is misleading at best. Experimental methods and their closest cousins were developed and refined because they generally do a good job providing answers to *one* kind of question. Specifically, these methods, under certain assumptions, give a good estimate of the effect that a potential causal variable has on one or more outcomes of interest. For example, an experiment could be conducted to assess the effect of the FrankenGene web-based tool (the potential causal variable) on introductory biology students' performance on an objective test and their reported interest in STEM majors and professions (the outcomes of interest). Because the outcomes of an educational product or program are frequently the things

*RCTs can be a terrific option for summative evaluation, especially for a relatively mature program or product.*

that matter most for an evaluative judgment, RCTs can be a terrific option for summative evaluation, especially for a relatively mature program or product.

In contrast, RCTs may have limited value when other evaluation purposes are of primary interest. Take formative evaluation. In principle, an RCT could be carried out to test the relative effectiveness of one way of implementing an REU versus another, but in general this would be overkill. The cost of the RCT would necessitate a narrower scope for the evaluation, thereby reducing the range and potential value of the formative evaluation. Now consider a proof of concept evaluation. For these, the key concern usually is simply whether things can be done in the new way being proposed, and evaluations should attend largely to the program's implementation rather than its outcomes.

Even when summative evaluation is of interest, it is not inherently a given that an experiment is called for. In transgenetics, it is not necessary to randomly assign dozens of pigs to receive the jellyfish gene while other pigs are randomly assigned not to receive the gene. Because pigs never glow in the dark otherwise, even a single demonstration with one pig is fairly compelling. However, for most of the kinds of projects that NSF funds in higher education, the effectiveness of the project is not so dramatic and clear. For an REU project designed to increase the number of STEM majors, for example, several considerations would make it difficult to see the effects with the "naked eye." Graduation with a STEM major is affected by numerous factors other than REU participation. Some of these factors are probably linked with the tendency to participate in an REU in the first place, making it hard to parse out the effects of the REU per se. In addition, students' enrollment in a STEM major can change over time, and the likelihood of entering or dropping out of a STEM major may vary across individuals; this kind of variability across time and across individuals makes it more difficult (if not impossible, without an adequate research design) to sort out the effects of REU participation with simple comparisons. Finally, because no magic bullet has yet been found, NSF and others still care about increases in STEM participation even if they are not so gigantic that they will be visible to the naked eye. Put differently, to use a phrasing drawn from the work of Donald Campbell and his colleagues (e.g., Campbell and Stanley, 1966; Shadish, Cook, and Campbell, 2002), experiments and their closest cousins are especially important when there are plausible alternative explanations of findings that otherwise might be interpreted as an effect of the intervention. Because the effects we can reasonably expect are not as dramatic as a glow-in-the-dark pig, and because many factors other than the NSF project affect outcomes, more effort is needed to get an accurate answer.

*Experiments and their closest cousins are especially important when there are plausible alternative explanations of findings.*

When alternative explanations are plausible, RCTs can be valuable. In a randomized experiment, individuals (or other units, such as classrooms)

are randomly assigned to one of two or more treatment conditions. For example, some introductory biology students could be assigned at random—essentially, with a flip of a coin or, more likely, by a random number table or a computerized equivalent—to use the FrankenGene web-based tool, while others are assigned at random not to use it. A well-conducted randomized experiment helps rule out internal validity threats, which are generic kinds of alternative explanations. Put differently, these validity threats are ways of accounting for the summative evaluation results, other than a conclusion such as "the project works." To understand the benefits of random assignment, consider first an alternative. Imagine that instead of random assignment in the FrankenGene example, the web-based tool was simply made available to students in introductory biology courses. In the absence of random assignment, one might compare the achievement of students who chose to use FrankenGene with those who did not (perhaps with careful measurement of how much each student used the tool). However, the validity threat called *selection* would apply. That is, any observed difference between those who did and those who did not use FrankenGene could easily have resulted, not from the effect of the web-based tool, but from preexisting differences on any of a number of variables. For example, the students who chose to use FrankenGene might initially have had greater interest in STEM, stronger motivation to do well in class, or a better work ethic. These and other confounds (i.e., factors unintentionally correlated with FrankenGene use) could easily obscure the true effect of the web-based tool.

An evaluator might try to account for confounds by measuring them and controlling for them statistically. In the FrankenGene example, the evaluator might measure students' preexisting interest in STEM disciplines. However, this approach assumes that the relevant initial differences are known and well measured, which will not necessarily be the case. In contrast, random assignment effectively takes care of selection problems (and, assuming the experiment is conducted successfully, other internal validity threats). If students are assigned to conditions at random, no systematic selection bias will exist. Because each student is equally likely to be assigned to the treatment and comparison groups, within statistical limits the two groups should not differ unless the treatment works. Moreover, any random differences between the two groups can be effectively accommodated with familiar hypothesis-testing statistics. The ability to estimate treatment effects without bias (i.e., without the intrusion of selection and other internal validity threats) provides the primary argument for randomized experiments.

*The ability to estimate treatment effects without bias provides the primary argument for randomized experiments.*

In short, advocates of RCTs have a point. RCTs are a potentially strong method for estimating the effects of a given intervention. They generally do well in terms of taking care of selection and other threats that could otherwise bias the estimate of the intervention's effects. Not surprisingly, however, RCTs also face challenges. Attrition may occur, and in ways that could bias the results. Contamination across conditions can occur, for instance, if students in the FrankenGene condition share both

*The circumstances that allow random assignment may be unusual, making it more tenuous to generalize the evaluation findings to other settings.*

excitement about the site and access information with students who are supposed to be in the comparison condition. In some cases, the circumstances that allow random assignment may be unusual, making it more tenuous to generalize the evaluation findings to other settings. In addition, if the relevant outcomes are not specified in advanced or not measured well, the results may be misleading.

In short, an argument can be made for the use of randomized experiments. However, the argument generally will not make sense unless the primary evaluation purpose is summative, and usually for a later stage project. Moreover, even for summative, late-stage evaluations, project evaluation planning should involve thoughtful judgment rather than reflex. The judgment process should consider the relevant tradeoffs, the plausibility of validity threats, and the relative desirability in the particular case of an RCT and alternative methods. Paraphrasing Eleanor Chelimsky, past Director of the Program Evaluation and Methodology Division of the U.S. Government Accountability Office, the only gold standard for evaluation is methodological appropriateness.

## Alternative, Related Methods

Even when summative evaluation is of interest, an RCT may not be the best choice for project evaluation. What are some of the better alternatives to consider? *Quasi-experiments* are approximations of randomized experiments. They also can be worth considering for summative evaluation—and, in the weaker forms, for formative evaluation. Unlike randomized experiments, by definition quasi-experiments lack random assignment to conditions. At the same time, like randomized experiments, quasi-experiments typically involve comparisons across two or more conditions, such as using or not using FrankenGene. Quasi-experiments may also include before-after comparisons, as when faculty members in a summer workshop are tested at the beginning and end of the workshop.

Quasi-experiments rather than randomized experiments may be chosen for a number of reasons. First, quasi-experiments may be feasible when random assignment is not for practical or ethical reasons. Second, *sometimes* quasi-experiments, even relatively simple ones, can give a compelling answer to the question of what effects the treatment of interest has. For example, if a project involves a relatively new intervention for developing faculty expertise in STEM instruction, a simple design measuring participating faculty members' knowledge before and after the three-week summer training workshop may well suffice. In this instance, there may be few if any plausible alternative explanations for a gain in the relevant knowledge. A third reason for using a quasi-experiment rather than an RCT is that the level of

confidence needed may be low. For instance, the summative evaluation for an early stage project does not require as high a level of confidence, so an RCT would be overkill. In contrast, a relatively modest quasi-experiment may suffice. Fourth, even for later stage projects, a more advanced quasi-experiment should do about as well as an RCT and may be better in terms of tradeoffs.

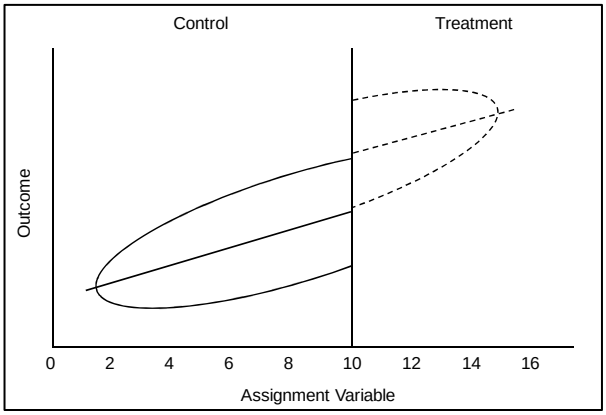
Numerous quasi-experimental designs and design variants exist. These are discussed in detail elsewhere (see Shadish, Cook, and Campbell, 2002; Mark and Reichardt, 2009). NSF project evaluators should have a working knowledge of the range of quasi-experimental designs. One relatively simple quasi-experiment, alluded to previously, can be called the *one-group, pretest-postest design*. In it, participants are measured on an outcome variable before and after the project activities. The evaluator's hope is that if the project is effective, outcome scores should improve, while scores will hold steady if the project has no effect. However, a variety of validity threats exist. These include *maturation*, the possibility that outcome scores will change because of ordinary maturational processes as the participants age. Another threat is *history*, the possibility that some event other than the project is responsible for any observed change. Because of maturation, history, and other potential threats,[4] you often would not get an accurate answer about the project's effect simply by measuring the relevant outcome variable both before and after. On the other hand, when validity threats are not plausible, or a confident answer is not required, the one-group pretest/posttest design often is a "best buy."

*Numerous quasi-experimental designs and design variants exist.*

"Stronger" quasi-experimental designs, such as the so-called regression-discontinuity design and complex interrupted time series designs, generally tend to rule out more validity threats than do "weaker" quasi-experiments such as the one-group, pretest-posttest design. It is these stronger quasi-experimental designs that were being referred to in previous mentions of the "closest cousins" of RCTs. It is not possible to describe here the full range of quasi-experimental designs and features. Rather, selected quasi-experimental designs and design elements are overviewed. In the case of an *interrupted time series design*, data on an outcome of interest are tracked repeatedly over time, well before an intervention is implemented and afterwards. For example, an evaluator might track the number and percent of minority students in STEM majors at a school with an LSAMP project. If a clear increase in these measures coincides with the onset of the LSAMP project, that would be consistent with the idea that the alliance is effective. This conclusion would be even stronger if the increase in minority participation in STEM professions occurs at the school with the LSAMP project but not at similar comparison schools in the same geographic region.

[4] Among the other reasons, in the language of Campbell and his colleagues, are the validity threats of history, maturation, statistical regression, testing, instrumentation, and attrition.

In the case of the *regression-discontinuity (R-D) design*, certain rare circumstances have to occur in order to use the design. However, if these circumstances occur, you can implement a very strong quasi-experimental design. In the R-D design, study participants all receive a score on an "assignment variable," or AV, and they are *assigned to groups based on this score*. Those who score above a specified cutoff value on the AV are assigned to one treatment group, while those who score below the cutoff value are assigned to the other group. For example, to encourage participation in a project for college faculty, the project might publicize the workshop as highly selective and an honor. Applicants might have to agree to participate in testing whether they are chosen or not, and to provide a portfolio and essay that project staff will score on a 100-point scale. In this example, the AV would be the score from the portfolio and essay grading. Faculty members with scores above the cutoff would be assigned to the workshop group, while those below the cutoff serve as a comparison group. Subsequently, all applicants, both workshop and comparison group members, would be tested on the outcome measure, perhaps by a web-based survey. In essence, statistical analysis of the R-D design involves testing whether the scores of those above the cutoff are elevated, relative to what would be expected given the pattern of scores below the cutoff. If this occurs, there rarely are any plausible threats to internal validity. Put informally, how likely is it that there would be a jump in scores on the outcome variable that coincides precisely with the cutoff on the AV, unless there really is a treatment effect? Analysis of the R-D design is more complex, but the underlying logic remains simple.




In another quasi-experimental design, the *pretest/posttest nonequivalent group design*, two groups (say, one participating in project activities and one not) are observed on both a pretest and a posttest. The design is not as strong generally as the R-D and interrupted time series designs, but often it will be more feasible. With this design, the researcher can use the pretest to try to take account of initial differences between the groups (i.e., the validity threat of selection). The basic logic is straightforward: The project's effect on the outcome measures is estimated in terms of how much more (or less) the project group gained on average than the comparison group. In fact, alternative analyses exist that imply somewhat different technical definitions of the project effect, but the fundamental logic remains the same. That is, measures of preproject differences are used to control for the threat of selection. A problem with this design is that it can be difficult to know that you've properly controlled for the differences between groups that would exist in the absence of the project (or, equivalently, if the project in fact made no difference). Consider the old expression, "The rich get richer." If the students who sign up for an REU project differ from comparison students

in their initial interest in STEM careers, it is possible that those differences would intensify over time even if there had been no REU, as the students get closer to graduation. Because different analysis approaches make slightly different assumptions about the pattern of change over time, it is advisable for a high-stakes summative evaluation using this design to employ more than one form of analysis to see if results are similar. An analysis approach that is increasingly recommended is *propensity score analysis*, which combines multiple preproject variables into a single variable to control for initial differences. (You might think of it as matching project and comparison group members, but on a composite of many variables, rather than on a single one—it's matching on steroids, so to speak).

Whether using an RCT or a quasi-experiment for the summative component of a design, certain issues arise that might not be familiar to investigators from some STEM disciplines. One is *statistical power*, which refers roughly to the likelihood that if a difference really exists between the groups, the study will be able to observe it. Metaphorically, statistical power is akin to whether the power of magnification of a microscope is sufficient to have a good chance of observing some microbe of interest if it actually is in the sample being examined. Projects often are small enough in size that statistical power is a problem. In such cases, sometimes power can be increased by combining across multiple cohorts. For instance, rather than treating each summer workshop in a TUES faculty development project as separate, the evaluation might employ analyses that combine across years.

*Statistical power is akin to whether the power of magnification of a microscope is sufficient to have a good chance of observing some microbe of interest if it actually is in the sample being examined.*

*Mediation* is another issue that can be addressed in later-stage summative evaluations. Mediators are shorter-term variables that change as a result of the project activities and that in turn result in change in longer-term outcomes. For example, for an LSAMP or other pipeline project, shorter-term changes in interest in STEM careers and in self-efficacy in STEM might be expected to occur first, with these changes mediating the effect of the project on STEM major and career choices. Metaphorically, meditational analysis involves seeing whether one domino (project activities) knocks over the next (interest, self-efficacy), with those in turn knocking down the later dominos (STEM major/career choices). Mediation can be tested with advanced statistical analyses, in essence testing whether the dominoes fell as expected. Qualitative approaches can also be employed, say by interviewing project participants about the changes they experience.

## Conclusions

Let us return to a question posed early in this chapter, "What does the 'best' or 'most rigorous' evaluation look like for a STEM education

project funded by NSF?" The short answer, again, is "It depends." A longer answer is that it depends on a number of considerations, including the relative priority of alternative evaluation purposes; the stage or maturity of the thing being evaluated; the information needs and potential uses of important evaluation consumers; the state of relevant knowledge; the evaluation requirements in the NSF proposal solicitation; the specifics of the project and what it is intended to do; the details of tradeoffs between breadth, rigor, cost, and time; how confident an answer is needed; and the method options that the project details make feasible or infeasible.

It is not possible to look at every combination of the factors that might influence the design of a project evaluation and then to describe the ideal evaluation. Nor would that be desirable—evaluation planning should not be a paint-by-number exercise. Fortunately, the proposal solicitation will usually provide a general framework, based on NSF staff members' understandings of the factors most relevant for that particular program. And the hope is that the material in this chapter and the examples throughout will help allow more thoughtful consideration of the options as you plan your project evaluation.

**References**


Campbell, D. T., & Stanley, J. C. (1996). *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Chen, H-t. (2004). *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness*. Thousand Oaks, CA: Sage.

Mark, M. M., & Reichardt, C. S. (2009). Quasi-experimentation. In *The SAGE Handbook of Applied Social Research Methods,* 2nd ed, edited by L. Bickman and D. Rog, 182-213. Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

## Appendix A

### Finding an Evaluator

There are many different sources for locating a project evaluator. The one that works best will depend on a number of factors including the home institution for the project, the nature of the project, and whether or not the principal investigator has some strong feeling about the type(s) of evaluation that are appropriate.

There are at least three avenues that can be pursued:

- If the project is being carried out at or near a college or university, a good starting point is likely to be at the college or university itself. Principal investigators can contact the department chairs from areas such as education, psychology, administration, or sociology and ask about the availability of staff skilled in project evaluation. In most cases, a few calls will yield several names.
- A second source for evaluation assistance comes from independent contractors. There are many highly trained personnel whose major income derives from providing evaluation services. Department chairs may well be cognizant of these individuals and requests to chairs for help might include suggestions for individuals they have worked with outside of the college or university. In addition, independent consultants can be identified from the phone book, from vendor lists kept by procurement offices in state departments of education and in local school systems, and even from resource databases kept by some private foundations, such as the Kellogg Foundation in Michigan.
- Finally, suggestions for evaluators can be obtained from calls to other researchers or perusal of research and evaluation reports. Western Michigan University also has a list of evaluators on their website at http://www.wmich.edu/evalctr. A strong personal recommendation and a discussion of an evaluator's strengths and weaknesses from someone who has worked with a specific evaluator is very useful when starting a new evaluation effort.

Although it may take a chain of telephone calls to get the list started, most principal investigators will ultimately find that they have several different sources of evaluation support from which to select. The critical task then becomes negotiating time, content, and, of course, money.

## Appendix B

## Glossary

**Accuracy:** The extent to which an evaluation is truthful or valid in what it says about a program, project, or material.

**Achievement:** Performance as determined by some type of assessment or testing.

**Affective:** Consists of emotions, feelings, and attitudes.

**Anonymity (provision for):** Evaluator action to ensure that the identity of subjects cannot be ascertained during the course of a study, in study reports, or in any other way.

**Assessment:** Often used as a synonym for evaluation. The term is sometimes recommended for restriction to processes that are focused on quantitative and/or testing approaches.

**Attitude:** A person's opinion about another person, thing, or state.

**Attrition:** Loss of subjects from the defined sample during the course of data collection.

**Audience(s):** Consumers of the evaluation; those who will or should read or hear of the evaluation, either during or at the end of the evaluation process. Includes those persons who will be guided by the evaluation in making decisions and all others who have a stake in the evaluation (see stakeholders).

**Authentic assessment:** Alternative to traditional testing that focuses on student skill in carrying out real-world tasks.

**Background:** Information that describes the project, including its goals, objectives, context, and stakeholders.

**Baseline:** Facts about the condition or performance of subjects prior to treatment or intervention.

**Behavioral objectives:** Measurable changes in behavior that are targeted by a project.

**Bias:** A point of view that inhibits objectivity.

**Case study:** An intensive, detailed description and analysis of a single project, program, or instructional material in the context of its environment.

**Categorical scale:** A scale that distinguishes among individuals by putting them into a limited number of groups or categories.

**Checklist approach:** The principal instrument for practical evaluation, especially for investigating the thoroughness of implementation.

**Client:** The person or group or agency that commissioned the evaluation.

**Coding:** To translate a given set of data or items into descriptive or analytic categories to be used for data labeling and retrieval.

**Cohort:** A term used to designate one group among many in a study. For example, "the first cohort" may be the first group to have participated in a training program.

**Component:** A physically or temporally discrete part of a whole. It is any segment that can be combined with others to make a whole.

**Conceptual scheme:** A set of concepts that generate hypotheses and simplify description, through the classification and categorization of phenomena, and the identification of relationships among them.

**Conclusions (of an evaluation):** Final judgments and recommendations.

**Content analysis:** A process using a parsimonious classification system to determine the characteristics of a body of material or practices.

**Context (of an evaluation):** The combination of factors accompanying the study that may have influenced its results, including geographic location, timing, political and social climate, economic conditions, and other relevant professional activities in progress at the same time.

**Continuous scale:** A scale containing a large, perhaps infinite, number of intervals. Units on a continuous scale do not have a minimum size but rather can be broken down into smaller and smaller parts. For example, grade point average (GPA) is measured on a continuous scale, a student can have a GPA or 3, 3.5, 3.51, etc. (See categorical scale.)

**Criterion, criteria:** A criterion (variable) is whatever is used to measure a successful or unsuccessful outcome, e.g., grade point average.

**Criterion-referenced test:** Test whose scores are interpreted by referral to well-defined domains of content or behaviors, rather than by referral to the performance of some comparable group of people.

**Cross-case analysis:** Grouping data from different persons to common questions or analyzing different perspectives on issues under study.

**Cross-sectional study:** A cross-section is a random sample of a population, and a cross-sectional study examines this sample at one point in time. Successive cross-sectional studies can be used as a substitute for a longitudinal study. For example, examining today's first year students and today's graduating seniors may enable the evaluator to infer that the college experience has produced or can be expected to accompany the difference between them. The cross-sectional study substitutes today's seniors for a population that cannot be studied until four years later.

**Data display:** A compact form of organizing the available information (for example, graphs, charts, matrices).

**Data reduction:** Process of selecting, focusing, simplifying, abstracting, and transforming data collected into written field notes or transcriptions.

**Delivery system:** The link between the product or service and the immediate consumer (the recipient population).

**Descriptive data:** Information and findings expressed in words, unlike statistical data, which are expressed in numbers.

**Design:** The process of stipulating the investigatory procedures to be followed in doing a specific evaluation.

**Dissemination:** The process of communicating information to specific audiences for the purpose of extending knowledge and, in some cases, with a view to modifying policies and practices.

**Document:** Any written or recorded material not specifically prepared for the evaluation.

**Effectiveness:** Refers to the worth of a project in achieving formative or summative objectives. "Success" is its rough equivalent.

**Elite interviewers:** Well-qualified and especially trained persons who can successfully interact with high-level interviewees and are knowledgeable about the issues included in the evaluation.

**Ethnography:** Descriptive anthropology. Ethnographic program evaluation methods often focus on a program's culture.

**Executive summary:** A nontechnical summary statement designed to provide a quick overview of the full-length report on which it is based.

**External evaluation:** Evaluation conducted by an evaluator outside the organization within which the project is housed.

**Field notes:** Observer's detailed description of what has been observed.

**Focus group:** A group selected for its relevance to an evaluation that is engaged by a trained facilitator in a series of discussions designed for sharing insights, ideas, and observations on a topic of concern to the evaluation.

**Formative evaluation:** Evaluation designed and used to improve an intervention, especially when it is still being developed.

**Goal:** A broad-based description of an intended outcome.

**Hypothesis testing:** The standard model of the classical approach to scientific research in which a hypothesis is formulated before the experiment to test its truth.

**Impact evaluation:** An evaluation focused on outcomes or payoff of a project.

**Implementation evaluation:** Assessing program delivery (a subset of formative evaluation).

**In-depth interview:** A guided conversation between a skilled interviewer and an interviewee that seeks to maximize opportunities for the expression of a respondent's feelings and ideas through the use of open-ended questions and a loosely structured interview guide.

**Informed consent:** Agreement by the participants in an evaluation to the use, in specified ways for stated purposes, of their names and/or confidential information they supplied.

**Instrument:** An assessment device (test, questionnaire, protocol, etc.) adopted, adapted, or constructed for the purpose of the evaluation.

**Internal evaluator:** A staff member or unit from the organization within which the project is housed.

**Inter-rater reliability:** A measure of the extent to which different raters score an event or response in the same way.

**Intervention:** Project feature or innovation subject to evaluation.

**Intra-case analysis:** Writing a case study for each person or unit studied.

**Key informant:** Person with background, knowledge, or special skills relevant to topics examined by the evaluation.

**Longitudinal study:** An investigation or study in which a particular individual or group of individuals is followed over a substantial period of time to discover changes that may be attributable to the influence of the treatment, or to maturation, or the environment. (See also cross-sectional study.)

**Matrix:** An arrangement of rows and columns used to display multi-dimensional information.

**Measurement:** Determination of the magnitude of a quantity.

**Meta-evaluation:** Evaluation of the merit of the evaluation itself.

**Mixed-method evaluation:** An evaluation for which the design includes the use of both quantitative and qualitative methods for data collection and data analysis.

**Moderator:** Focus group leader; often called a facilitator.

**Nonparticipant observer:** A person whose role is clearly defined to project participants and project personnel as an outside observer or onlooker.

**Norm-referenced tests:** Tests that measure the relative performance of the individual or group by comparison with the performance of other individuals or groups taking the same test.

**Objective:** A specific description of an intended outcome.

**Observation:** The process of direct sensory inspection involving trained observers.

**Ordered data:** Nonnumeric data in ordered categories (for example, students' performance categorized as excellent, good, adequate, and poor).

**Outcome:** Post-treatment or post-intervention effects.

**Paradigm:** A general conception, model, or "worldview" that may be influential in shaping the development of a discipline or subdiscipline (for example, "the classical, positivist social science paradigm in evaluation").

**Participants:** Those individuals who are directly involved in a project.

**Participant observer:** An evaluator who participates in the project (as participant or staff) in order to gain a fuller understanding of the setting and issues.

**Performance evaluation:** A method of assessing what skills students or other project participants have acquired by examining how they accomplish complex tasks or the quality of the products they have created (e.g., poetry, artwork).

**Population:** All persons in a particular group.

**Prompt:** Reminder used by interviewers to obtain complete answers.

**Purposive sampling:** Creating samples by selecting information-rich cases from which one can learn a great deal about issues of central importance to the purpose of the evaluation.

**Qualitative evaluation:** The approach to evaluation that is primarily descriptive and interpretative.

**Quantitative evaluation:** The approach to evaluation involving the use of numerical measurement and data analysis based on statistical methods.

**Random sampling:** Drawing a number of items of any sort from a larger group or population so that every individual item has a specified probability of being chosen.

**Recommendations:** Suggestions for specific actions derived from evidence-based conclusions.

**Sample:** A part of a population.

**Secondary data analysis:** A reanalysis of data using the same or other appropriate procedures to verify the accuracy of the results of the initial analysis or for answering different questions.

**Self-administered instrument:** A questionnaire or report completed by a study participant without the assistance of an interviewer.

**Stakeholder:** One who has credibility, power, or other capital invested in a project and thus can be held to be to some degree at risk with it.

**Standardized tests:** Tests that have standardized instructions for administration, use, scoring, and interpretation with standard printed forms and content. They are usually norm-referenced tests but can also be criterion referenced.

**Strategy:** A systematic plan of action to reach predefined goals.

**Structured interview:** An interview in which the interviewer asks questions from a detailed guide that contains the questions to be asked and the specific areas for probing.

**Summary:** A short restatement of the main points of a report.

**Summative evaluation:** Evaluation designed to present conclusions about the merit or worth of an intervention and recommendations about whether it should be retained, altered, or eliminated.

**Transportable:** An intervention that can be replicated in a different site.

**Triangulation:** In an evaluation, an attempt to get corroboration on a phenomenon or measurement by approaching it by several (three or more) independent routes. This effort provides confirmatory measurement.

**Utility:** The extent to which an evaluation produces and disseminates reports that inform relevant audiences and have beneficial impact on their work.

**Utilization of (evaluations):** Use and impact are terms used as substitutes for utilization. Sometimes seen as the equivalent of implementation, but this applies only to evaluations that contain recommendations.

**Validity:** The soundness of the inferences made from the results of a data-gathering process.

**Verification:** Revisiting the data as many times as necessary to cross-check or confirm the conclusions that were drawn.

## Appendix C.
## Bibliographies

### Annotated Bibliography on Readings in Evaluation

In this section, we summarize some evaluation references that readers of this Handbook might want to consult for additional information. We believe the selected references will be especially useful for NSF principal investigators and project directors. Additional references, without annotation, are presented in the next section.

American Evaluation Association. *Guiding Principles for Evaluators*. Revisions reflected herein ratified by the AEA membership, July 2004**.** Last retrieved from http://www.eval.org/Publications/GuidingPrinciplesPrintable.asp on December 23, 2010.

In 1994, the American Evaluation Association established a set of principles to guide the practice of evaluation. The *Guiding Principles for Evaluators* can help you identify the basic ethics to expect from an evaluator. They include:

1. Systematic Inquiry – Evaluators conduct systematic, data-based inquires about whatever is being evaluated.

2. Competence – Evaluators provide competent performance to stakeholders.

3. Integrity/honesty – Evaluators ensure the honesty and integrity of the entire evaluation process.

4. Respect for people – Evaluators respect the security, dignity, and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.

5. Responsibilities for general and public welfare – Evaluators clarify and take into account the diversity of interests and values that may be related to the general and public welfare.

Callow-Heusser, C., Chapman, H., & Torres, R. (2005). *Evidence: An Essential Tool.* Arlington, VA: National Science Foundation. Last retrieved from http://nsf.gov/pubs/2005/nsf0531/nsf0531.pdf on December 23, 2010.

Written for NSF's Math and Science Partnership (MSP) Program, this document provides a guiding framework for conducting evaluation in an

R&D environment. The document discusses the need for high-quality evidence of effectiveness and efficiency. It presents the DIO Cycle of Evidence as a guiding framework for planning, gathering, and using evidence at three stages: design, implementation, and outcome assessment. The document also discusses the relationship between this framework and other frameworks for evaluation.

Donaldson, S.I., Christin, C.A., & Mark, M.M. (2009). What Counts as Credible Evidence in Applied Research and Evaluation Practice? Thousand Oaks, CA: Sage Publications

Building on a symposium held at Claremont University, this volume brings together the ideas of an international group of researchers and evaluators regarding the issue of what counts as credible evidence in different evaluation contexts. The volume explores both experimental and non-experimental approaches, considering theoretical, methodological, political, ethnic, and pragmatic concerns.

Joint Committee on Standards for Educational Evaluation. (2010). *The Program Evaluation Standards.* 3rd Ed. Thousand Oaks, CA: Sage Publications.

This second version of standards for program evaluation provides guidance to evaluators and evaluation managers on how to judge an evaluation's quality, considering both methodological and ethical issues. The standards fall into four categories: utility, feasibility, propriety, and accuracy. Illustrative case studies are presented to help describe practices that meet the standards, as well as those that fall short.

Patton, M.Q.(2008) *Utilization-Focused Evaluation.* 4th Ed. Thousand Oaks, CA: Sage Publications.

In a book that combines both the theoretical and the practical, Patton examines how and why to conduct evaluations. The author discusses strategies for increasing the probability that an evaluation will be useful and be used, starting from the beginning of evaluation planning. This edition covers a range of issues from identifying the primary users of an evaluation to focusing the evaluation, making methods decisions, analyzing data, and presenting findings. Both formative and summative evaluation are discussed, along with the different roles that may be played by the evaluator in different situations.

Patton, M.Q. (2011). *Developmental Evaluation Applying Complexity Concepts to Enhance Innovation and Use*. New York, NY: The Guilford Press.

This book discusses developmental evaluation, an approach in which the evaluator is part of the project's design team. Rather than thinking of evaluation as a two-stage process involving formative and summative

phases, developmental evaluation recognizes the organic nature of complex projects providing ongoing data on emergent and changing activities. Situated in systems theory, developmental evaluation encourages an approach which is reflective, adaptive, and utilization focused.

Rossi, P.H., Lispey, M.W., & Freeman, H.E. 2004. Evaluation: A Systematic Approach. 7th Ed. Thousand Oaks, CA: Sage Publications.

Provides an overview of evaluation and the different types of activities that evaluation may include, with chapters on assessing (a) program theory, (b) measuring and monitoring program outcomes, (c) assessing program impact using randomized field experiments, (d) assessing program impact using alternative designs, and (e) detecting, interpreting, and analyzing program effects, and measuring efficiency.

W.K. Kellogg Foundation. (2004). *Logic Model Development Guide.* Battle Creek, MI: Last retrieved from http://www.wkkf.org/knowledge-center/resources/2010/Logic-Model-Development-Guide.aspx on December 23, 2010.

*The Logic Model Development Guide* provides an easy-to-read description of the logic model and how it can be applied to a variety of different evaluation situations. This is an excellent "starter" document for those who may be new to the logic model and its application.

Wholey, J., Hatry, H., & Newcomber, K. (Eds.). (2004) *Handbook of Practical Program Evaluation.* 2nd Ed. San Francisco, CA: John Wiley & Sons.

This book provides an overview of how to do more effective and useful evaluation, starting with design and working through making sure results are used. It is oriented toward developing evaluations that can be used for program improvement. Written by a variety of evaluators from a range of content fields, the authors offer advice on evaluation procedures, including ones that may not be ideal but are still likely to provide useful and reasonably reliable information at an affordable cost. As stated in the Preface, the philosophy underlying the book is "It's better to be roughly right than to be precisely ignorant."

## Appendix C.
## Bibliographies

### Annotated Bibliography on Readings On Cultural Context, Cultural Competence, and Culturally Responsive Evaluation

Compiled by Rodney Hopson, Tanya Brown, Liya Aklilu, Maurice Samuels, Lutheria Peters, and Kien Lee

Association for the Study and Development of Community. (2001) *Principles for Evaluating Comprehensive Community Initiatives*. Gaithersburg, MD: ASDC.

This report was produced by the ASDC on behalf of the national funding collaborative on violence prevention. The primary audiences for the document are evaluators and practitioners involved in comprehensive community initiatives (CCIs). The document lists 27 principles that were created to guide evaluators in engaging evaluation participants in a responsive manner and build upon the efficacy of community interventions and development. The principles are organized according to nine major themes: (a) engagement of practitioners, community participants, funders, and other stakeholders; (b) role of evaluator; (c) implementation of the evaluation process; (d) issues of power; (e) identification and definition of outcomes; (f) multiple levels of change; (g) attribution of results to the CCI; (h) utilization; and (i) standards of evaluation.

Greene, J. (2006). Evaluation, Democracy, and Social Change. In *Handbook of Evaluation: Policies, Programs, and Practice*, edited by I.F. Shaw, J.C. Greene, & M. Mark. London:Sage.

This chapter situates larger conceptualizations of democracy, equality, and justice in evaluation by focusing on the macro positioning of evaluation in society (i.e., issues related to which purposes and whose interests evaluation should serve) and the micro character of evaluation practice (i.e., the relationships evaluators establish with others in a given context and the processes and interactions that enact these relationships). The chapter presents and delineates a historical landscape of democratically oriented evaluation beginning with democratic evaluation and deliberative democratic evaluation to a discussion of participatory, critical, and culturally and contextually responsive evaluation. The author situates culturally and contextually responsive evaluation as an ideologically oriented contemporary evaluation that attends to culture, race, and ethnicity issues relevant both to racial and ethnic groups in the United States and to indigenous peoples in North America and the Pacific.

Greene, J., Millet, R., & Hopson, R. (2004). Evaluation as Democratizing Practice. In *Putting Evaluation to Work for Foundations and Grantees,* edited by M. Braverman, N. Constantine, and J.K. Slater, 96-118. San Francisco: Jossey-Bass.

The authors make a case for an educative and democratizing vision for evaluation in philanthropy. That is, evaluation is more than method and design—it is "inherently and fundamentally, a matter of politics and values." Evaluation is politically located within social contexts. Significance then lies not in asking what approaches should be taken when entering into a given context, but which "political positions and whose values should be advanced in the social practice of evaluation." The authors position evaluation within the tradition of policy education and claim that foundations can assert leadership in legitimizing and promulgating efforts and social change. Authors provide a discussion on how to enact the educative and democratizing vision of evaluation in practice by outlining three interconnected major principles. The principles are elaborated with illustrative guidelines, strategies, and examples from practice.

Guzmán, B. (2003). Examining the Role of Cultural Competency in Program Evaluation: Visions for New Millennium Evaluators. In *Evaluating Social Programs and Problems: Visions for the New Millennium,* edited by S. Donaldson and M. Scriven. Mahwah, NJ: Lawrence Erlbaum.

The chapter, written as proceedings to the Stauffer Symposium on Applied Psychology at the Claremont Colleges, identifies the need for increased cultural sensitivity in the increasingly multicultural and multiethnic American society. The chapter defines four characteristics that define culture: (a) culture as an abstract, human idea; (b) culture as a context of or setting within which behavior occurs, is shaped, and is transformed; (c) culture as containing values, beliefs, attitudes, and languages that have emerged as adaptations; and (d) culture as important to be passed on intergenerationally. The author asserts that building culturally competent evaluators is more complicated than developing cookie-cutter or cookbook approaches to integrating issues for diverse cultural groups but involves considerable effort and engagement between evaluators, participants, and clients to ensure the evaluation process incorporates cultural norms and adaptations throughout.

Hood, S., Hopson, R., & Frierson, H. (2005). *The Role of Culture and Cultural Context in Evaluation: A Mandate for Inclusion, the Discovery of Truth, and Understanding in Evaluative Theory and Practice.* Greenwich, CT: Information Age.

This book argues for and provides examples of evaluators who consider how to be responsive to cultural context and how to adopt strategies that are congruent with cultural understandings in evaluation theory, history,

and practice. The authors encourage the evaluation community to develop efforts to address issues related to training culturally competent evaluators, designing culturally competent evaluations, and enhancing the usefulness of these efforts. In providing a broad array of issues related to culturally responsive evaluation, chapters in the book consistently express the need for change in the traditional ways of practicing educational evaluation from drawing on experiences in indigenous and other sociocultural contexts to developing theory-driven approaches that are unique to communities of color. In doing so, the authors use persuasive communication, narrative, and other strategies of illustrating the possibilities for privileging the impact of culture in evaluation. Ultimately, through the use of philosophical, historical, theoretical, and practical illustrations in the United States, the authors offer hope to redress the shortcomings of evaluation theory and practice that omit matters related to culture.

Hood, S. (2001). *Nobody Knows My Name: In Praise of African American Evaluators Who Were Responsive*. New Directions in Evaluation, 92: 31-44. San Francisco: Jossey-Bass.

In this paper, the author demonstrates how the work of early African American scholars, whose work went largely unnoticed in the fields of educational evaluation, builds upon contemporary notions of responsive and culturally responsive evaluation. The author outlines how these evaluators demonstrated the practices of responsive evaluation decades prior to formal evaluation theories and practice of responsive evaluation. He notes that these proponents were not only responsive, but embodied critical practices of a *culturally* responsive evaluation practice. By engaging the *shared lived experiences* of key stakeholders in the evaluation context, early African American evaluators were able to better depict the key concerns of the stakeholders. He further argues that their work supports the significance of a diversified field of evaluators. Based upon these contributions, the author identifies an imperative within the field to cultivate more scholars of color.

Hopson, R. (2001). Global and Local Conversations on Culture, Diversity, and Social Justice in Evaluation: Issues to Consider in a 9/11 Era. *American Journal of Evaluation,* 20(3): 375-380.

This paper contemplates the future of the profession of evaluation by exploring cross-cultural concerns within the field. The author illustrates how valuable lessons regarding evaluation practice and theory may be found across continents in his discussion of developments in the African Evaluation Association, and the Namibian Evaluation Network (NEN) in particular. The author identifies the politics of transferring evaluation standards and guidelines across continents without critical reflection of the relevance of the standards in a foreign context. He notes the comparatively insufficient attention to issues related to cultural competence within North America, and suggests reconstructing methodologies and paradigms that historically have served to cripple or

debilitate underserved or marginalized communities. He discusses how African evaluators have worked responsively to address stakeholder concerns in practice and in their construction of evaluation guidelines aimed to protect the communities they serve.

Hopson, R. (2003). *Overview of Multicultural and Culturally Competent Program Evaluation: Issues, Challenges & Opportunities.* Oakland, CA: California Endowment.

Written for the California Endowment, this paper details the history and impact of efforts addressing multiculturalism and cultural competence in the field of evaluation. The author regards recent attention to multicultural/culturally competent evaluation as demarcating a paradigm shift in the field, as it exposes the "epistemological ethnocentrism" that has privileged dominant world views and the values of the white middle class. He provides a historical overview of multicultural/ culturally competent evaluation, identifies key professional organizations and meetings that served as catalysts for promoting these perspectives, and synthesizes his arguments by outlining five basic tenets of multicultural/culturally competent evaluation. The author closes the paper with a discussion of the implications for moving the evaluation field forward.

House, R. E. (2001). *Responsive Evaluation and Its Influence on Deliberative Democratic Evaluation.* New Directions for Evaluation, 92: 23-30.

The author reviews the evolution of Stake's responsive evaluation and the manner that its key elements have changed the structure of evaluation practice. He begins by highlighting how Stake resolved many of the fissures within evaluation practice in the late 1960s and early 1970s. Stakes' argument for practice to inform later theory drastically shifted evaluation approach. The author outlines the features of responsive evaluation, critiques elements of his structure, and offers new perspectives (i.e., social justice in evaluation). He notes that the strength of responsive evaluation is that it helped break the intellectual stranglehold that single-method approaches had on evaluation at one time and legitimated multiple avenues for conducting evaluations.

House, E.R. (1993). Evaluation in Multicultural Societies. In *Professional Evaluation: Social Impact and Political Consequences*, edited by E. R. House, 141-162. Newbury Park, CA: Sage Publications.

The ninth chapter in the book focuses on the issues that pluralist nation-states face in incorporating different cultural groups and how these same issues translate into problems for evaluation, namely what criteria is needed to employ in evaluating programs, which stakeholders to include in the evaluation, and how to balance the various ethnic interests in

drawing conclusions. In an early attempt to think about notions of culturally responsive evaluation, the chapter foreshadows how stakeholder approaches in evaluation should more closely and deliberately "search out and define the views and interests of these minority cultures if they are stakeholders in the program being evaluated." In addition, by combining philosophical elements of social justice theory and minority rights and interests, the chapter provides definitions of nationalism and ethnicity and unpacks how Canada and the United States manifest as unique multicultural societies.

Kirkhart, K. E. (1995). Seeking Multicultural Validity: A Postcard from the Road. *Evaluation Practice*, 16(1): 1-12.

In her American Evaluation Association presidential address, the author proposes the construct *multicultural validity* to achieve two aims: (a) to organize concerns about pluralism and diversity in evaluation, and (b) to reflect on the cultural boundedness of evaluation work. Her paper underscores an ethical imperative to address issues of culture and context when practicing in the evaluation field. She defines multicultural validity as the ability to capture diverse perspectives within an evaluation context accurately, soundly, and appropriately. She argues that multicultural validity is a necessary prerequisite to social justice and builds upon the traditional understandings and uses of validity. She unpacks this argument by a discussion of multiculturalism and the construct of culture, outlining the multiple dimensions and purposes of validity, and listing threats to multicultural validity.

Madison, A-M. (1992). *Minority Issues in Program Evaluation.* New Directions for Evaluation, 53.

This seminal special issue helped to lay the early foundation of thinking on the topic of cultural issues in evaluation; specifically, the issue aims to "begin discussion of some minority concerns about the impact of cultural dominance on definitions of social goals and on the measurement of their outcomes in a culturally diverse society, and about the political consequences for minorities of cultural dominance in the selection of evaluation methods." Each chapter presents an evaluation issue (e.g., impacting policy on minority youth and adults, exploring potential for developing programs and evaluations that incorporate racial and ethnic minorities into evaluation experience, defining the limitations of current evaluation models and current techniques for understanding the impact of social policy on the lives of minority groups).

Sengupta S., Hopson, R., & Thompson-Robingson, M. (2004). *Cultural Context in Evaluation: An Overview.* New Directions for Evaluation, 102: 5-21.

The authors make a call to the evaluation profession to establish policies and practice guidelines addressing cultural competence in the evaluation. They refer to developments that have occurred within other disciplines and the relative lag within the evaluation field. The authors argue however, that a deeper exploration of culture is critically important within evaluation because of the critical role that policy, program, and service delivery play in operationalizing cultural contexts. The concept of values is presented as a common thread between evaluation and culture. The authors note that serious consideration of values in evaluation exposes the degree to which some evaluation methods have failed to address the values operating within a given evaluation context. They encourage the use of the term "cultural competence" within the evaluation field, and provide extended discussion of its meaning within the field. Lastly, the authors call for the development of multicultural and multifaceted evaluators and identify steps taking place in that direction.

Thomas, V. G., & Stevens, F. (2004). *Co-Constructing a Contextually Responsive Evaluation Framework: The Talent Development Model of School Reform.* New Directions for Evaluation, 101.

This special edited issue presents the Talent Development (TD) evaluation framework, an approach for evaluating urban school reform interventions. Rooted in responsive, participatory, empowerment, and culturally competent approaches in evaluation, the Howard University Center for Research on the Education of Students Placed at Risk (CRESPAR) Talent Development evaluation approach includes themes that emphasize inclusiveness, cooperation, and usefulness of individuals being served by evaluations. The issue includes chapters that introduce the Talent Development Model of School Reform and illustrative case examples of the TD evaluation framework in practice, including commentaries that assess the extent to which the framework demonstrates a coherent evaluation approach and that explore the utility of critical race theory in the educational evaluation process specific to the TD evaluative paradigm.

Thompson-Robinson, M., Hopson, R., & SenGupta, S. (2004). In *Search of Cultural Competence in Evaluation: Toward Principles and Practices*. New Directions for Evaluation, 102.

This special edited issue addresses a number of important questions as they relate to culture in evaluation. Specifically how does culture matter in evaluation theory and practice? How does attention to cultural issues make for better evaluation practice? What is the value-addedness of cultural competence in evaluation? The issue includes an overview of culture, cultural competence, and culturally competent evaluation and includes case studies on the implementation of culturally competent

evaluation in school settings, tribal settings, programs that cater to HIV/AIDS populations, and other diverse settings. Additionally, contributors present lessons learned from their experiences both locally and globally, and offer recommendations for implementing culturally competent evaluations in general that are systematic and deliberate in program and institutional planning and development.

## Appendix C.
## Bibliographies

### Other Recommended Reading


Boykin, L.L. (1957). Let's Eliminate the Confusion: What is Evaluation? *Educational Administration and Supervision,* 43 (2): 115-121.

Debus, M. (1995). *Methodological Review: A Handbook for Excellence in Focus Group Research*. Washington, DC: Academy for Educational Development.

Denzin, N.K., & Lincoln, Y.S. (Eds.). (1994). *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage.

Erlandson, D.A., Harris, E.L., Skipper, B.L., & Allen, D. (1993). *Doing Naturalist Inquiry: A Guide to Methods*. Newbury Park, CA: Sage.

Fox, S. (2000). An Effective School Evaluation and Training Program. In *The Cultural Context of Educational Evaluation: The Role of Minority Evaluation Professionals* (NSF 01-43). Arlington, VA: National Science Foundation, Directorate for Education and Human Resources.

Frierson, H.T. (2000). The Need for the Participation of Minority Professionals in Educational Evaluation. In *The Cultural Context of Educational Evaluation: The Role of Minority Evaluation Professionals* (NSF 01-43). Arlington, VA: National Science Foundation, Directorate for Education and Human Resources.

Greenbaum, T.L. (1993). *The Handbook of Focus Group Research*. New York: Lexington Books.

Hart, D. (1994). *Authentic Assessment: A Handbook for Educators*. Menlo Park, CA: Addison-Wesley.

Herman, J.L., & Winters, L. (1992). *Tracking Your School's Success: A Guide to Sensible Evaluation*. Newbury Park, CA: Corwin Press.

Hood, S. (2001). Nobody Knows My Name: In Praise of African American Evaluators Who Were Responsive. In *Responsive Evaluation: Roots and Wings*, edited by J. Greene & T. Abma. New Directions for Program Evaluation. San Francisco, CA: Jossey-Bass.

Hood, S. (2000). A New Look at an Old Question. In *The Cultural Context of Educational Evaluation: The Role of Minority Evaluation Professionals* (NSF 01-43). Arlington, VA: National

Science Foundation, Directorate for Education and Human Resources.

Hughes, G. (2000). Evaluation of Educational Achievement of Underrepresented Minorities: Assessing Correlates of Student Academic Achievement. In *The Cultural Context of Educational Evaluation: The Role of Minority Evaluation Professionals* (NSF 01-43). Arlington, VA: National Science Foundation, Directorate for Education and Human Resources.

Hymes, D.L., Chafin, A.E., & Gondor, R. (1991). *The Changing Face of Testing and Assessment: Problems and Solutions.* Arlington, VA: American Association of School Administrators.

Krueger, R.A. (1988). *Focus Groups: A Practical Guide for Applied Research*. Newbury Park, CA: Sage.

LeCompte, M.D., Millroy, W.L., & Preissle, J. (Eds.). (1992). *The Handbook of Qualitative Research in Education.* San Diego, CA: Academic Press.

Merton, R.K., Fiske, M., & Kendall, P.L. (1990*). The Focused Interview: A Manual of Problems and Procedures.* 2nd Ed. New York: The Free Press.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, CA: Sage.

Morgan, D.L. (Ed.). (1993). *Successful Focus Groups: Advancing the State of the Ar*t. Newbury Park, CA: Sage.

Morse, J.M. (Ed.). (1994). *Critical Issues in Qualitative Research Methods*. Thousand Oaks, CA: Sage.

National Science Foundation. (2001). *The Cultural Context of Educational Evaluations: The Role of Minority Evaluation Professionals.* Workshop Proceedings. June 1-2, 2000.

Perrone, V. (Ed.). (1991). *Expanding Student Assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Reich, R.B. (1991). *The Work of Nations*. New York: Alfred A. Knopf.

Rodriquez, C. (2000). Assessing Underrepresented Science and Mathematics Students: Issues and Myths. In *The Cultural Context of Educational Evaluation: The Role of Minority Evaluation Professionals* (NSF 01-43). Arlington, VA: National Science Foundation, Directorate for Education and Human Resources.

Sanders, J.R. (2000). *Evaluating School Programs.* 2nd Ed. Thousand Oaks, CA: Corwin Press.

Schatzman, L., & Strauss, A.L. (1973). *Field Research*. Englewood Cliffs, NJ: Prentice-Hall.

Seidman, I.E. (1991). *Interviewing as Qualitative Research: A Guide for Researchers in Education and Social Sciences.* New York: Teachers College Press.

Smith, M.L. (1986). The Whole Is Greater: Combining Qualitative and Quantitative Approaches in Evaluation Studies. In *Naturalistic Evaluation*, edited by D. Williams. New Directions for Program Evaluation, Vol. 30. San Francisco, CA: Jossey-Bass, Inc.

Stake, R. (1972). *Program Evaluation, Particularly Responsive Evaluation.* ERIC Document ED 075-187. Last retrieved from http://eric.ed.gov/PDFS/ED075487.pdf on December 23, 2010

Stewart, D.W., and Shamdasani, P.N. (1990). *Focus Groups: Theory and Practice.* Newbury Park, CA: Sage.

U.S. General Accounting Office (GAO). (1990). *Case Study Evaluations.* Paper 10.1.9. Washington, DC: GAO.

Weiss, R.S. (1994). *Learning From Strangers: The Art and Method of Qualitative Interview Studies*. New York: Free Press.

Wiggins, G. (1989). A True Test: Toward More Authentic and Equitable Assessment. *Phi Delta Kappan*, May, 703-704.

Wiggins, G. (1989). Teaching to the (Authentic) Test. *Educational Leadership*, 46, 45.

**NATIONAL SCIENCE FOUNDATION**
ARLINGTON, VA 22230

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE $300

**RETURN THIS COVER SHEET TO ROOM P35 IF YOU DO NOT WISH TO RECEIVE THIS MATERIAL , OR IF CHANGE OF ADDRESS IS NEEDED , INDICATE CHANGE INCLUDING ZIP CODE ON THE LABEL (DO NOT REMOVE LABEL).**