DOCUMENT RESUME

ED 468 812 SE 066 794

AUTHOR	Frechtling, Joy
TITLE	The 2002 User-Friendly Handbook for Project Evaluation.
INSTITUTION	Westat, Rockville, MD.
SPONS AGENCY	National Science Foundation, Arlington, VA. Directorate for Education and Human Resources.
REPORT NO	NSF-02-057
PUB DATE	2002-01-00
NOTE	94p.; Special section by Henry Frierson, Stafford Hood, and Gerunda Hughes.
CONTRACT	REC-99-12175
AVAILABLE FROM	National Science Foundation, Directorate for Education and Human Resources, 4201 Wilson Blvd., Arlington, VA 22230. Tel: 301-947-2722. For full text: http://www.nsf.gov.
PUB TYPE	Guides - Non-Classroom (055)
EDRS PRICE	EDRS Price MF01/PC04 Plus Postage.
DESCRIPTORS	*Evaluators; *Evaluation Methods; Evaluative Thinking; General Education; Program Evaluation
IDENTIFIERS	*National Science Foundation

ABSTRACT

This handbook was developed to provide managers working with the National Science Foundation (NSF) with a basic guide for the evaluation of NSF's educational programs. It is aimed at people who need to learn more about what evaluation can do and how to do an evaluation rather than those who already have a solid base of experience in the field. This handbook builds on firmly established principles, blending technical knowledge and common sense to meet the special needs of NSF and its stakeholders. Quantitative and qualitative evaluation methods are discussed, 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 program managers will increase their understanding of the evaluation process and NSF's requirements for evaluation as well as gain knowledge that will help them communicate with evaluators and manage the actual evaluation. Several NSF program areas were selected to provide concrete examples of the evaluation issues discussed. The handbook is divided into four major sections: (1) "Evaluation and Types of Evaluation"; (2) "The Steps in Doing an Evaluation"; (3) "An Overview of Quantitative and Qualitative Data Collection Methods"; and (4) "Strategies That Address Culturally Responsive Evaluation." A glossary of commonly used terms, references for additional readings, and an appendix that presents some tips for finding an evaluator are also included. (MM)

Directorate for Education
and Human Resources

Division of Research,
Evaluation and Communication
National Science Foundation

The 2002 User-Friendly Handbook for Project Evaluation

THE 2002 USER-FRIENDLY HANDBOOK FOR PROJECT EVALUATION

The 2002 User Friendly Handbook for Project Evaluation

Prepared under Contract
REC 99-12175

by

Joy Frechtling
Westat

with a special section by

Henry Frierson
Stafford Hood
Gerunda Hughes

Conrad Katzenmeyer
Program Officer and COTR
Division of Research, Evaluation and Communication
National Science Foundation

January 2002

The National Science Foundation
Directorate for Education & Human Resources
Division of Research, Evaluation, and Communication

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 managers working with the National Science Foundation (NSF) with a basic guide for the evaluation of NSF's educational programs. It is aimed at people who need to learn more about both what evaluation can do and how to do 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 program managers 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 manage the actual evaluation.

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* and the 1997 *User-Friendly Handbook for Mixed-Method Evaluations*) 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, several NSF program areas have been selected to provide concrete examples of the evaluation issues discussed. The Handbook is divided into four major sections:

- Evaluation and types of evaluation
- The steps in doing an evaluation
- An overview of quantitative and qualitative data collection methods
- Strategies that address culturally responsive evaluation

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 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.

Section 1

EVALUATION AND TYPES OF EVALUATION

1. REASONS FOR CONDUCTING EVALUATIONS

The notion of evaluation has been around 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. In addition to its long history, evaluation also has varied definitions and may mean different things to different people. Evaluation can be seen as synonymous with tests, descriptions, documents, or even management. Many definitions have been developed, but a comprehensive definition 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 deciding a course of action.

Why should NSF grantees do evaluation? There are two very important answers to this question. First and foremost, evaluation provides information to help improve the project. Information on whether goals are being met and on how different aspects of a project are working are essential to a continuous improvement process. 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 are among the most useful outcomes of the assessment enterprise.

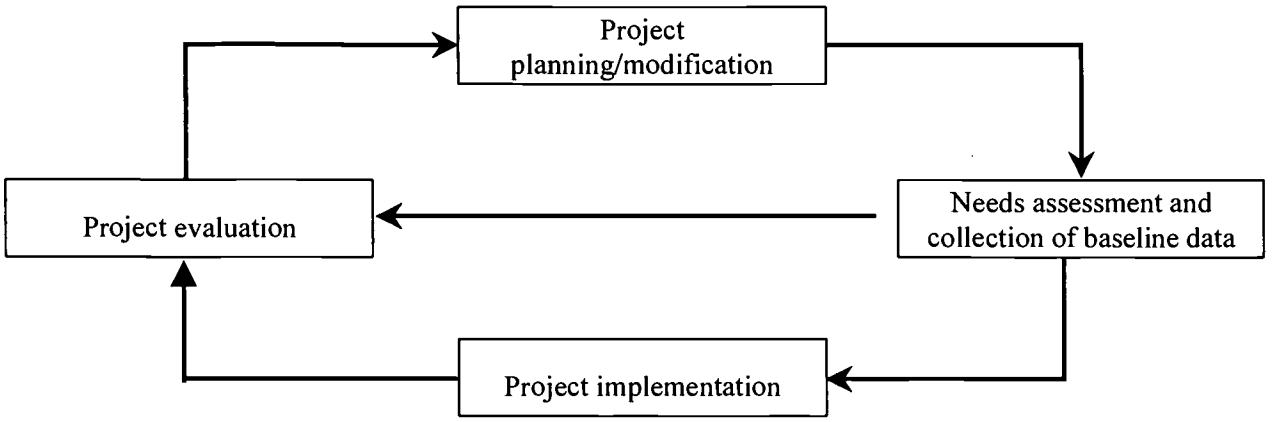
Evaluation provides information to help improve a project.

Over the years, evaluation has frequently been viewed as an adversarial process. Its main use has been to provide a "thumbs-up" or "thumbs-down" about a program or project. In this role, 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.

Evaluations need not be conducted in an adversarial mode.

The current view of evaluation stresses the inherent interrelationships between evaluation and program implementation. Evaluation is not separate from, or added to, a project, but rather is part of it from the beginning. Planning, evaluation, and implementation are all parts of a whole, and they work best when they work together. Exhibit 1 shows the interaction between evaluation and other aspects of your NSF project.

Exhibit 1.—The project development/evaluation cycle



Evaluation provides information for communicating to a variety of stakeholders.

Second, evaluation provides information for communicating to a variety of stakeholders. It allows projects to better tell their story and prove their worth. It also gives managers the data they need to report "up the line," to inform senior decisionmakers about the outcomes of their investments. This notion of reporting on the outcomes of federal investments has received increased emphasis over the last several years with the establishment of the Government Performance and Results Act (GPRA). GPRA requires federal agencies to report annually on the accomplishments of their funded efforts. This requirement includes establishing broad goals or strategic outcomes, performance outcomes, and performance indicators against which progress will be assessed. GPRA goes beyond counts of who is funded or who is served, placing the focus instead on results or impacts of the federal investment. In response, NSF has chosen to focus on three general strategic outcomes:

- Developing a diverse internationally competitive and globally engaged workforce of scientists, engineers, and well-prepared citizens;
- Enabling discoveries across the frontiers of science and engineering connected to learning, innovations, and service to society; and
- Providing broadly accessible, state-of-the-art information bases and shared research and education tools.

Projects will be asked to provide data on their accomplishments in these areas, as relevant. Detailed requirements for the information to be provided have been developed on a program-by-program basis.

[1] NSF, FY 2002 GPRA Performance Plan, April 19, 2001, p. 2.

Project directors should keep GPRA and these strategic outcomes in mind in developing plans for project evaluation (more information on NSF's approach to GPRA can be found at www.nsf.gov/od/gpra/start.htm).

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" describes 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's start by defining terms and showing how they relate.

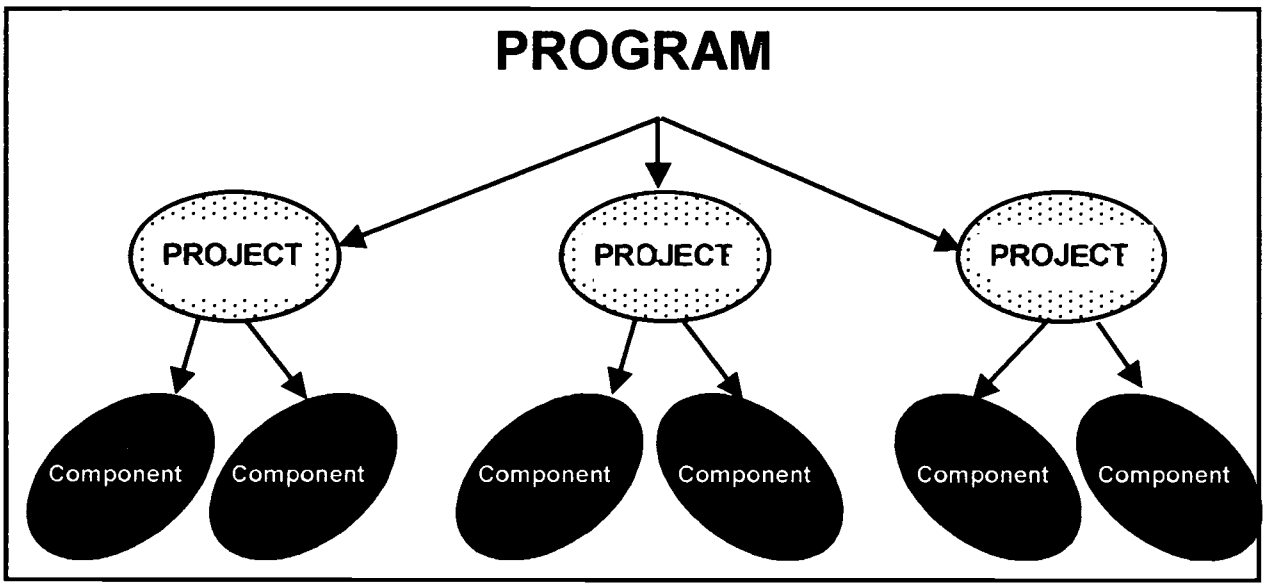
A **program** is a coordinated approach to exploring a specific area related to NSF's mission of strengthening science, mathematics, and technology. A **project** is a particular investigative or developmental activity funded by that program. NSF initiates a program on the assumption that an agency goal (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's 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, and 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.

Project evaluations might also include examination of specific critical 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 looks to see the extent to which its goals have been met (these goals are a subset of the overall project goals), and to clarify the extent to which the component contributes to the success or failure of the overall project.

Exhibit 2.—Levels of evaluation

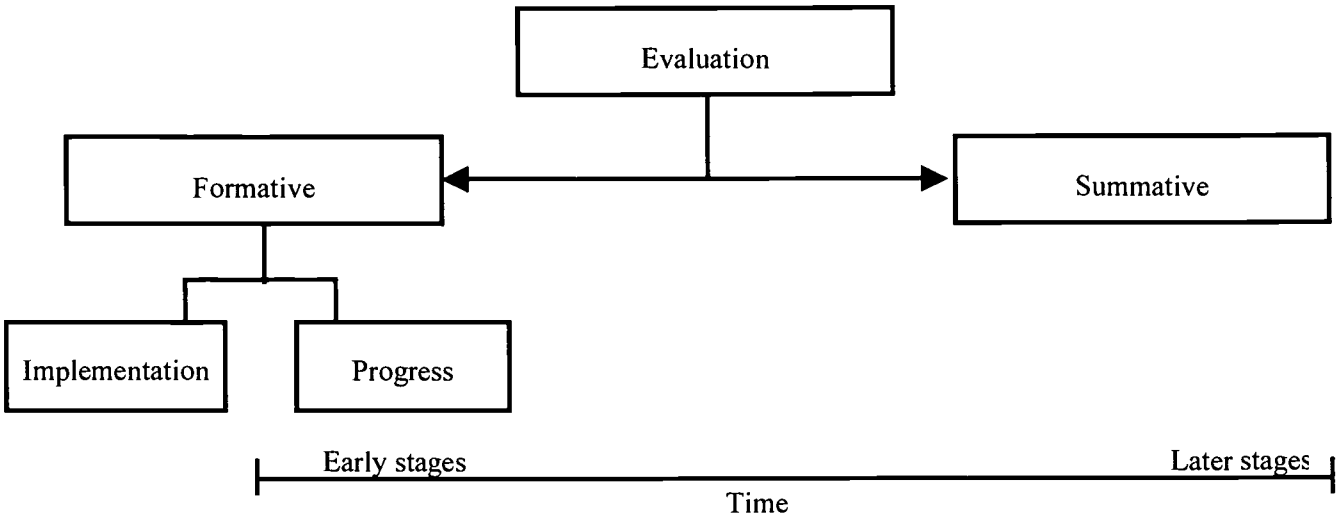


The information in this Handbook has been developed primarily for the use of project directors and principal investigators, although project evaluators may also find it useful. Our aim is to provide tools that will help those responsible for the examination of individual projects gain the most from their 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 kinds or stages of evaluation—formative evaluation and summative evaluation. The purpose of a formative evaluation is to assess initial and ongoing project activities. 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 throughout the life of the project. Its intent is to assess ongoing project activities and provide information to monitor and improve the project. It is done at several points in the developmental life of a project and its activities. 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," may occur once or several times during the life of the program. The underlying principle is that before you can evaluate the outcomes or impact of a program, you must make sure the program and its components are really operating and, if they, are operating according to the proposed plan or description.

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

A series of implementation questions guides an implementation evaluation. For example, questions that might be posed for the NSF Louis Stokes Alliances for Minority Participation (LSAMP) 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?
- Do the activities and strategies match those described in the plan? 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 project implementation 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 goals of the program and the project. It involves collecting information to learn whether or not the benchmarks of participant progress were met and to point out unexpected developments. Progress evaluation collects information to determine what the impact of the activities and strategies is on participants, curriculum, or institutions at various stages of the intervention. By measuring progress, program staff can eliminate the risk of waiting until participants have experienced the entire program to assess likely outcomes. 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 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 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 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 (sometimes referred to as impact or outcome evaluation) frequently addresses many of the same questions as a progress evaluation, but it takes place after the project has been established and the timeframe posited for change has occurred. A summative evaluation of an LSAMP project might address these basic questions:

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

- To what extent does the project meet the stated goals for change or impact?
- 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 schools participants attend? Are there any changes in courses? Are there any impacts of the LSAMP program on overall 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 program's cost?
- Can the program be sustained?
- Is the program replicable and transportable?

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

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 decisionmaking. 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; and discontinue.

In most situations, especially high-stakes situations or situations that are politically charged, it is important to have an external evaluator who is seen as objective and unbiased. Appendix A provides some tips for finding an evaluator. If this is not possible, it is better to have an internal evaluation than none at all. One compromise between the external and the internal model 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 program 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 Gathering Activities

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

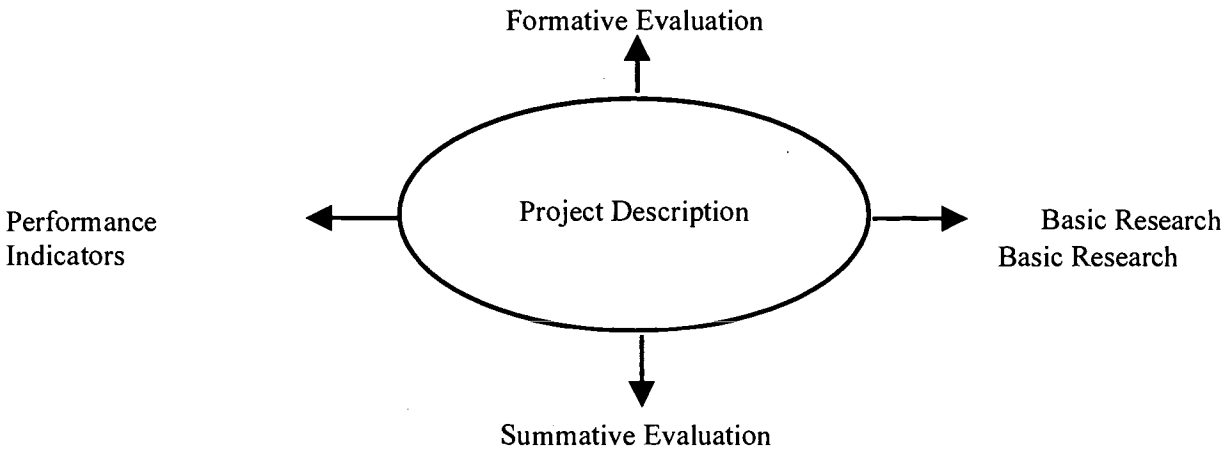
It is useful to understand how evaluation complements, but may differ from, other types of data collection activities that provide information on accountability for an NSF-funded project. Exhibit 4 shows various types of data collection activities, each of which provides somewhat different information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

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 gathering 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. In addition, some programs have added program-specific modules aimed at collecting tailored data elements.

Exhibit 4.—Types of data gathering 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 more indepth data collection activities, are intended to support decisionmaking, and are more costly.

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. Going beyond descriptive statistics, 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 program expenditures and activity levels. In fact, the term "performance" underscores the underlying purpose of indicator systems, i.e., to examine a program's accomplishments 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 indepth exploration of issues, but differ along other dimensions. Instead of being intended for decisionmaking, research efforts typically are designed to explore conceptual models and alternative explanations for observed relationships.

Summary

The goal of evaluation is 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. In these days of reform, educators are continually 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 decisionmaking 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.

Section II THE STEPS IN DOING AN EVALUATION

3. THE EVALUATION PROCESS— GETTING STARTED

In the preceding chapter, we outlined the types of evaluations that should be considered for 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 evaluation studies.

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

- Develop a conceptual model of the program and identify key evaluation points
- Develop evaluation questions and define measurable outcomes
- Develop an evaluation design
- Collect data
- Analyze data
- Provide information to interested audiences

Getting started right can have a major impact on the progress and utility of the evaluation all along the way. However, all six phases are critical to providing useful information. 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 contribute to the decisionmaking process.

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.

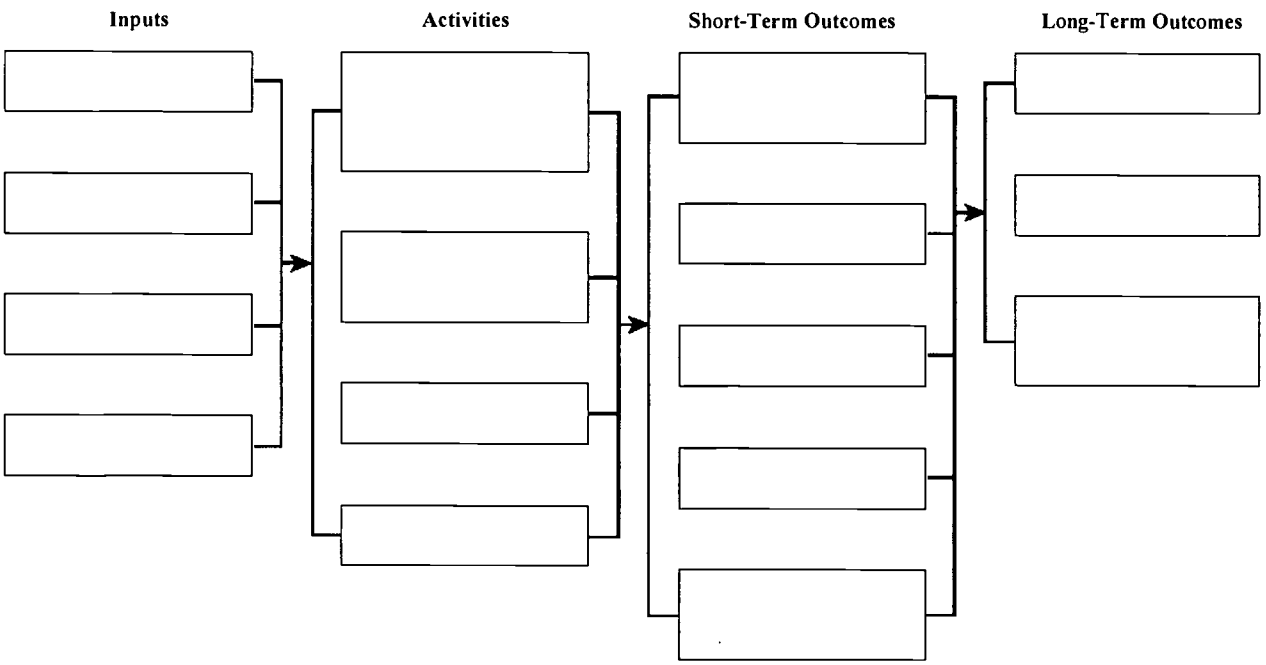
Develop a Conceptual Model of the Project and Identify Key Evaluation Points

Every proposed evaluation should start with a conceptual model to which the design is applied. 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."[2] The model describes the pieces of the project and expected connections among them. A typical model has four categories of project elements that are connected by directional arrows. These elements are:

- Project inputs
- Activities
- Short-term outcomes
- Long-term outcomes

Exhibit 5.—Logic model



[2] There are several different ways to show a logic model. The model presented here is one that has been useful to the author.

Project inputs are the various funding sources and resource streams that provide support to the project. Activities are the services, materials, and actions that characterize the project's thrusts. Short-term impacts are immediate results of these activities. Long-term outcomes are the broader and more enduring impacts on the system. These impacts will reflect NSF's strategic outcomes discussed on page 4. A logic model identifies these program elements and shows expected connections among them. PIs and PDs may find this model useful not only for evaluation but also for program management. It provides a framework for monitoring the flow of work and checking whether required activities are being put in place.

The first step in doing an evaluation is to describe the project in terms of the logic model.

- One set of inputs is the funds that NSF provides. Other inputs may come from other federal funding sources, local funding sources, partnerships, and in-kind contributions.

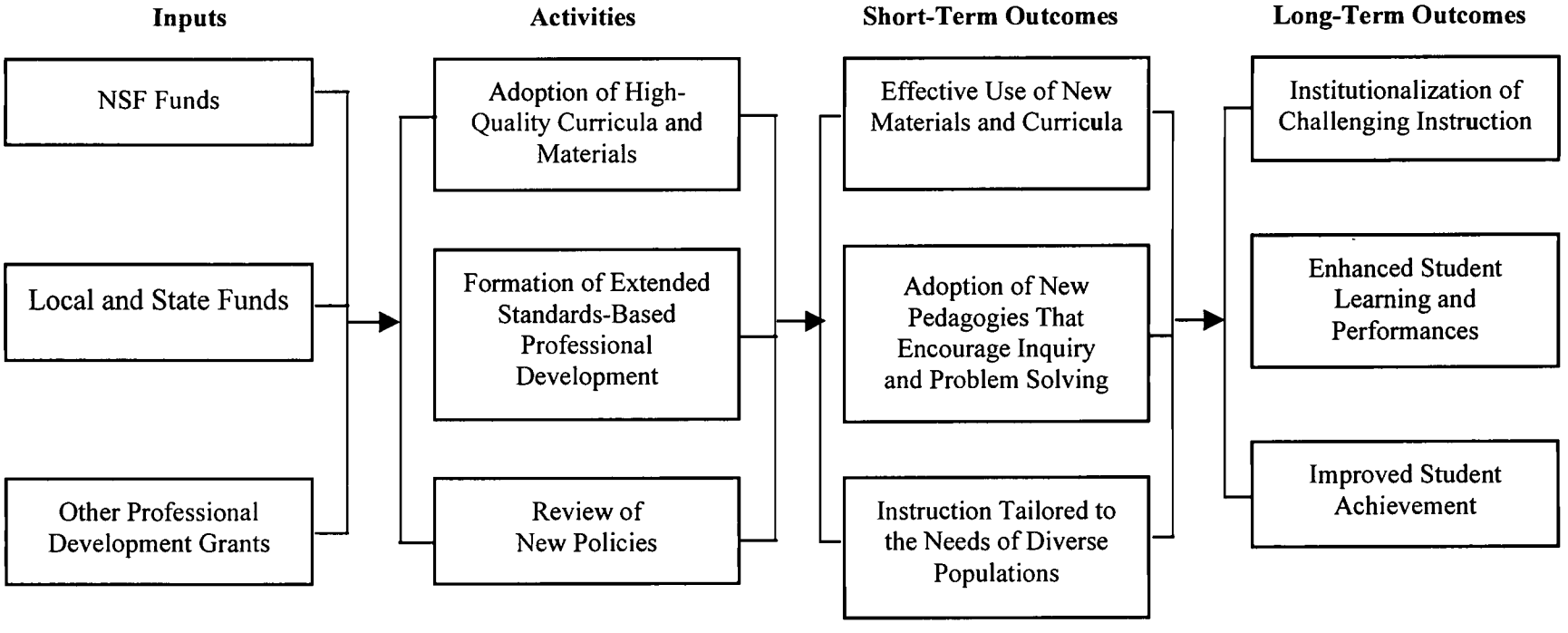
- The activities depend on the focus of the project. Potential activities include the development of curricula and materials, provision of professional development, infrastructure development, research experiences, mentoring by a senior scientist, or public outreach, alone or in combinations.

- Short-term outcomes in a variety of shapes and sizes. One type of outcome is sometimes called an "output." An output is an accounting of the numbers of people, products, or institutions reached. For example, an output of a professional development program for teachers could be "200 teachers trained." The output of a research program could be "17 students received mentoring from NSF scientists." The other type of outcome looks at short-term changes that result from the experience. Such an outcome might be "reported sense of renewal" for a teacher given professional development support or "an impact on choice of major" for an undergraduate receiving a research experience.

- Long-term outcomes are the changes that might not be expected to emerge until some time after the experience with the project. To continue with the examples provided above, a long-term outcome of professional development could be "changes in instructional practice reflective of a standards-based approach." For the undergraduate student, "selecting a career in NSF-related research activity" would be a comparable outcome.

The logic model shows a process that flows from inputs to long-term outcomes. 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

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



and then determining critical conditions or events that will need to be established before these outcomes might be expected to occur. Exhibit 6 shows a preliminary conceptual model for one of NSF's major professional development programs, Local Systemic Change Initiatives (LSCs) projects.

Under "inputs," we have listed three streams of funding:

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

For "activities," we have highlighted:

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

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 or students from diverse populations

Finally, over time, the LSCs should result in:

- Consistently challenging instruction for all students
- Enhanced student learning and performance
- Higher scores on assessments of student achievement

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 is an area that should have been addressed during the project's planning phase, and determining expected timeframes should be a revisiting of decisions rather than 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. Finally, the model can be used to

identify critical achievements as indicated by the logic model and critical timeframes that need to met. These provide the starting point for the next step, developing the evaluation questions.

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 may 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.

Although, in most cases, key stakeholders will share a number of information needs (in a professional development program 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 7 presents a shell for organizing your approach to identifying stakeholders and their specific needs or interests.

Exhibit 7.—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

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

These latter two points require some additional explanation. First, 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 thing, however, is determining a shared definition of what is meant and what will be accepted as credible evidence of project success. Exhibit 8 illustrates the steps to translating a general goal into a measurable objective.

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 1 to 5 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? First, for very large samples, "effect size" is frequently used as a second standard against which to measure the importance of an outcome. Using this approach, the change is measured against the standard deviation, and only those significant outcomes that result in a change that exceed one-third of a standard deviation are considered meaningful. Second, 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.

Exhibit 8.—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. Once you have completed the above steps, go back to #3 and write the next objective. Continue with steps 4, and 5, and 6.

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 given 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 LSC 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 5 to 10 percent of project cost for the evaluation.

A general guideline is to allocate 5 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:

- Selecting a methodological approach and data collection instruments
- Determining who will be studied and when

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-method" 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. There are a number of factors that need to be considered in reaching a decision regarding the methodologies that will be used. These include the questions being addressed, the timeframe 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.

Determining Who Will be Studied and When

Developing a design also requires considering factors such as sampling, use of comparison groups, timing, sequencing, and frequency of data collection.

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, i.e., 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's 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 9 describes each type of bias and suggests some simple ways of minimizing them.

Exhibit 9.—Three types of errors and their remedies

Type	Cause	Remedies
Sampling Error	Using a sample, not the entire population to be studied.	Larger samples—these 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 non-respondents 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 interviews, 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, 1990). 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.

Comparison Groups. 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. In classical evaluation design, this problem of attribution is addressed by creating treatment and control or comparison groups and randomly assigning the potential pool of participants to these varying conditions. In the ideal world, project evaluators would like to be able to adopt this same approach and examine program impacts under well-controlled experimental conditions. Unfortunately, in most real-world applications and most NSF projects, these conditions simply cannot be created.

There are two basic problems: first, there is self-selection. Teachers, students, and faculty participate in NSF efforts because they choose to, by and large. While there may be circumstances under which a participant is encouraged or even coerced into participating, that is likely to be the exception. Thus, there is reason to believe that those who volunteer or seek out programs are different from those who don't. Second, it is frequently difficult to identify a valid comparison group and obtain its cooperation with study efforts. The more elaborate and potentially intrusive the evaluation, the more difficult the task.

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

There is no perfect way to solve the problem, but in designing an evaluation it is important to address, rather than ignore, the attribution question. Sometimes this is possible by drawing a comparison group from a waiting list (when one exists) and comparing those who participated with those who self-selected but applied too late. Assuming that the groups are found to be equivalent on critical variables that might be associated with the outcome of interest, it is possible to relate differences to differences in program experiences.

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. 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.

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.

Timing, Sequencing, Frequency of Data Collection, and Cost. 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-method 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 strongest 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.

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

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 may become 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.

Once decisions are reached regarding the actual aspects of your evaluation design, it is useful to summarize these decisions in a design matrix. Exhibit 10 presents the shell for each matrix using the Minority Research Fellowship Program 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 10a.—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 10b.—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

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

Patton, M.Q. (1990). *Qualitative Evaluation and Research Methods*, 2nd Ed. Newbury Park, CA: Sage.

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

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:

- Data collection
- Data analysis
- Reporting
- Dissemination

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

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

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.

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. 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. Informed 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 has a set of Guiding Principles for Evaluators (AEA, 1995) 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 skills. 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. 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 Standards for Educational Evaluation, 1994).

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. This can mean sending surveys out two to three times or rescheduling interviews or observations on multiple occasions. An ambitious rule of thumb for surveys is to try to gather data from at least 80 percent of those sampled. 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, these should be noted and the impact on the generalizability of findings noted.

It is important to follow up with individuals who are non-responsive to the initial contact to try to get them to participate.

Finally, the data should be gathered, causing 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 on the part of the respondent to pull the

data together and develop summary statistics—the data collector may need to work from raw data, applications, and 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 the checking of 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. Recently, computerized systems for quantitative analysis have been developed and are becoming more widely used to manage large sets of narrative data. These provide support to the analyst

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

and a way of managing large amounts of data that are typically collected (but do not eliminate the need for careful analysis and decisionmaking on the part of the evaluator.) Two popular programs are Ethnograph and Nu*Dist.

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 address these further 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—i.e., 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 of these results emerge from the data; they are ones that were not anticipated or looked for. 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 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, or even structural equation modeling.

The final task is to choose the analyses to be presented, 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 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 typically include six major sections:

- Background
- Evaluation study questions
- Evaluation procedures

- Data analysis
- 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

The 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 gathering 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 will be 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; and
- 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 gathering approaches, the respondents, and the data collection schedule.

Data Analysis

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, 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.

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 4 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 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, which requires special formatting for reading or downloading from a web site. 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. 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. Today 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 11.

Exhibit 11.—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

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
- 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 three 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.

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. (1995). *Guiding Principles for Evaluators.* New Directions for Program Evaluation, No. 66. San Francisco, CA: Jossey-Bass.

Joint Committee on the Standards for Educational Evaluation (1994). *The Program Evaluation Standards*. Thousand Oaks, CA: Sage Publication, Inc.

Section III

AN OVERVIEW OF QUANTITATIVE AND QUALITATIVE DATA COLLECTION METHODS

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 science enrichment 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 special science 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 survey respondents may not understand the meaning of questions to which they respond, and because people's recall of events is often faulty. 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 about the respective merits of the two approaches, largely because of different views about the nature of knowledge and how knowledge is best acquired. 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 indepth 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, the 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 decisionmakers in the educational community where the project is located. In most cases, decisionmakers at the national level tend to 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 poor tool for assessing true student learning.

Staff Skills

Qualitative methods, including indepth interviewing, observations, and the use of focus groups, require good staff skills and considerable supervision 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 surveys, however, usually require more skilled personnel to design the instruments and to manage data collection and 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 short survey based on a small number of cases (25-50) and 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 respondents; while this method might provide more "interesting" data, those data would be primarily useful for generating new hypotheses to be tested by more appropriate qualitative or quantitative methods. To obtain robust findings, the cost of data collection is bound to be high regardless of method.

Time Constraints

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.

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 survey requires considerable time to create and pretest questions and to obtain high response rates. 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.

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-Method Approach

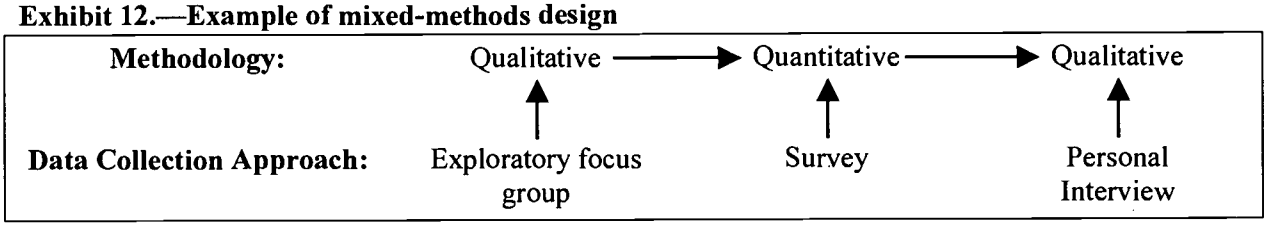
We feel that a strong case can be made for including qualitative elements in the great majority of evaluations of NSF projects. Most of the programs sponsored by NSF are not targeted to participants in a carefully

controlled and restrictive environment, but rather to those in a complex social environment that has a bearing on the success of the project. 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. A multimethod 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-method 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 12).

Exhibit 12.—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 in Altshuld and Witkin (2000).

But 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 indepth 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.

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.

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, indepth interviews, focus groups, observations, and tests. We also cover 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) web site (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 1 to 4 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 exploring the utility of survey 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. Even the traditional approach of mailing surveys for self-guided response has been supplemented by using facsimile for delivery and return.

Selecting the best method for collecting surveys requires weighing a number of factors. These included the complexity of questions, resources available, the project schedule, 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. However, at this time, unless the survey is fairly simple (no skip patterns,

limited use of matrices), the technology needed to develop such surveys can require a significant resource investment. As new tools are developed for commercial use, this problem should diminish.

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 indepth 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 13 shows the advantages and disadvantages of surveys.

Exhibit 13.—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 followup 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 indepth 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 indepth exploration of a more limited set of questions. Indepth interviews also encourage capturing respondents' perceptions in their own words, a very desirable strategy in qualitative data collection. This allows the evaluator to present the meaningfulness of the experience from the respondent's perspective. Indepth interviews are conducted with individuals or a small group of individuals.

When to Use Interviews

Interviews can be used at any stage of the evaluation process. Indepth 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 indepth interviews are particularly appropriate include situations involving complex subject matter, detailed information, high-status respondents, and highly sensitive subject matter. Exhibit 14 shows the advantages and disadvantages of interviews.

Exhibit 14.—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

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 8 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.

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 indepth 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 indepth interviews share many characteristics, they should not be used interchangeably. Factors to consider when choosing between focus groups and indepth interviews are displayed in Exhibit 15.

Observations

Observational techniques are methods by which an individual or individuals gather firsthand data on programs, 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.

Exhibit 15.—Which to use: Focus groups or indepth 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.
Logistics geographically	an acceptable number of target respondents can be assembled in one location.	respondents are dispersed or not easily assembled for other reasons.
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.

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. The technique would be especially useful in directly examining teaching methods employed by the faculty in their own classes after program participation. Exhibit 16 shows the advantages and disadvantages of observations.

Exhibit 16.—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 includes standards for performance at varying levels of proficiency, typically a three- or four-point scale ranging from below basic to advanced performance.

Criticisms of traditional, short-answer, norm-referenced tests have become widespread. These criticisms 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 decisionmaking. 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. The new tests, called performance assessments, 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.

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 17 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 the following web sites:

- http://www.wcer.wisc.edu/nise/Publications/Briefs/Vol_1_No_2/
- http://www.wcer.wisc.edu/nise/Publications/Research_Monographs/vol6.pdf
- http://www.wcer.wisc.edu/nise/Publications/Research_Monographs/vol118.pdf

Exhibit 17.—Advantages and disadvantages of tests

The advantages and disadvantage of tests depend largely on the typ 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 18 shows the advantages and disadvantages of document studies.

Exhibit 18.—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 offer expertise beyond the evaluation team. They are also very useful for assisting with the evaluation of curricula and other educational materials. 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 19 shows the advantages and disadvantages of key informants.

Exhibit 19.—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 principal investigator 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, 1989). Exhibit 20 shows the advantages and disadvantages of case studies.

Exhibit 20.—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 is likely to provide the most useful information.

References

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. (1990). *Qualitative Evaluation and Research Method*, 2nd Ed. Newbury Park, CA: Sage.

Yin, R.K. (1989). *Case Study Research: Design and Method*. Newbury Park, CA: Sage.

Section IV

STRATEGIES THAT ADDRESS CULTURALLY RESPONSIVE EVALUATION

7. A GUIDE TO CONDUCTING CULTURALLY RESPONSIVE EVALUATIONS

Henry T. Frierson, Stafford Hood, and Gerunda B. Hughes

Culture is a cumulative body of learned and shared behavior, values, customs, and beliefs common to a particular group or society. In essence, culture makes us who we are.

In doing project evaluation, it is also important to consider 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.

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

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 experience and understandings to the evaluation tasks at hand.

Why should a project director be concerned with the cultural context of a program undergoing evaluation? Simply put, as American society becomes increasingly diverse racially, ethnically, and linguistically, it is important that program designers, implementers, and evaluators understand the cultural contexts in which these programs operate. To ignore the reality of the existence of the influence of culture and to be unresponsive to the needs of the target population is to put the program in danger of being ineffective and to put the evaluation in danger of being seriously flawed.

Being sensitive and responsive to the culture of the participants and the cultural environment in which the programs exists is a process that should be an important component of program evaluation. Fortunately, cultural responsiveness as it relates to evaluation is gaining recognition

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

as a critical feature of the evaluation process. This is particularly true for programs in which the participants' culture is acknowledged to have a major impact on program outcomes.

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.

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

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. The responsibility that educational evaluators have is to recognize their own personal cultural preferences and to make a conscious effort to restrict any undue influence they might have on the work.

The second argument, that educational evaluations should not be in the business of *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. The field of educational evaluation has advanced over the past three decades, through its recognition of the role that fullness of description plays in a comprehensive evaluation process (e.g., Stake, 1967). In fact, it is becoming increasingly recognized 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.

This chapter discusses strategies that have been found to be useful in conducting culturally responsive evaluation and to identify areas where further help is needed. We examine the role of culturally responsive evaluation at each of the critical phases of the evaluation process, showing how its principles can be applied to enhance good inquiry.

Preparing for the Evaluation

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. At the very least, the evaluator or evaluation team should be fully aware of and responsive to the participants' and stakeholders' culture, particularly as it relates to and influences the program.

Multiethnic evaluation teams increase the chances of really hearing the voices of underrepresented students.

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 who 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, 2001).

Engaging Stakeholders

When designing an evaluation that seeks to be culturally responsive, considerable attention must be given to the identification of the stakeholders. Stakeholders play a critical role in all evaluations, especially culturally responsive ones, providing sound advice from the beginning (framing questions) to the end (disseminating the evaluation results). It is important to develop a stakeholder group representative of the populations the project serves, assuring 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. Students, for example, may qualify for consideration, as might their parents or care givers. When targeting an evaluation toward program improvement and decisionmakers' needs, it is easy to overlook the critical roles that students and parents might play in an educational evaluation.

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

In individual projects such as the Louis Stokes Alliance for Minority Participation and the Alliance for Graduate Education for the Professoriate, if participants' and stakeholders' perceptions and views are not taken into account from a cultural perspective, the evaluation may prove 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. It is important that all key voices are accurately heard and listened to. If they are not, the entire evaluation process may be limited in its accuracy and opportunities for meaningful program improvement drastically reduced.

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 program evaluation is designed to answer two basic questions: (1) Is the project being conducted as planned and is progress being made toward meeting its goals? and (2) 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 measure and describe program operations in order to "inform" project staff (and stakeholders) about the status of the program. Summative evaluations, on the other hand, reveal whether and to what extent the program achieved its goals and objectives.

Culturally responsive progress evaluations examine connections through culturally sensitive lenses.

Process evaluations examine the connections between and among program activities. Culturally responsive process evaluations examine those connections through culturally sensitive lenses. Careful documentation of the implementation of program 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 program 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 program timeline and the results of the process evaluation.

Summative evaluations provide information about program effectiveness. Culturally responsive summative evaluations examine the direct effects of the program implementation on the participants and attempt to explain the results within the context of the program. 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 program 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 significant 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 answers. Further, framing evaluative questions is *not* easily accomplished. In a culturally responsive evaluation, the questions will have been carefully considered not only by the evaluator and project staff, but 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 an agreed-upon list of questions has been articulated to the satisfaction of the evaluation team and stakeholders, an epistemological task of great import comes to the fore, but again, it is not an easy task. They must ask, "What will we accept as evidence when we seek answers to our evaluative questions?" This, too, should be decided *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. A democratic process also increases the likelihood that evaluative efforts will have all voices represented.

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.

There are a number of different evaluation designs that can be used to organize the processes of data collection and analysis and subsequently answer the evaluation questions. The evaluation design that you use does not necessarily need to be elaborate. It just needs to be appropriate for what you want to do.

As stated earlier, most comprehensive evaluation designs have both a qualitative and a quantitative component. Each component provides data in a format that is different from the other, but that can also be complementary to the other.

In addition, 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). Furthermore, when comparison or control groups are incorporated into the pre-test/post-test design, evaluators are able to determine to what extent some aspect of participants' behavior changed relative to where it would have been had they not been subject to the project intervention(s).

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. While it is preferable to use instruments that have some history, that have been tried out and have established validity and reliability, previous use 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.

Collecting the Data

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

Culturally responsive evaluation makes substantial use of qualitative evaluation techniques. One of the tenets of qualitative methodology is that the individual who is collecting the data is the instrument. With that in mind, an instrument (or individual) that is an improper measure 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 program 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.

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.

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

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 program 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 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. 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, using 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 the spirit of openness and self-improvement, the results for collecting culturally responsive evaluative data can be considerable. Similar improvements in data collection and interpretation through observation can be achieved through intensive training and mentoring. Although the authors commend such training, in-service training is not the preferred solution. Greater and longer lasting improvements in the collection of culturally responsive evaluative data and in the conduct of program evaluations can be realized principally by recruiting evaluation data collectors and analyzers who already possess a shared lived experience with those who are being evaluated.

Analyzing the Data

One may conduct appropriate statistical techniques, such as analyses of variance, and examine test score distributions without much concern for the cultural context in which the data were collected, although that may actually be somewhat shortsighted. But the analysis of interview data and the interpretation of descriptions of behavior related to programs undergoing evaluation cannot be achieved without 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 the creation of review panels principally comprising representatives from stakeholder groups to examine evaluative findings gathered by the principal evaluator and/or an evaluation team. When stakeholder groups composed of separate panels of parents, students, and community representatives, for example, review evaluative findings, the meaning of evaluative data is frequently fresh, and is not always aligned with confirming interpretations. Again, the results of the deliberations of review panels will not lend themselves necessarily to simple, easy answers. Our contention, however, is that they will more accurately reflect the complexity of the cultural context in which the data were gathered.

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

Disaggregation of collected data is a procedure that warrants increased attention. Disaggregation of data sets is highly recommended because evaluative findings that dwell exclusively on whole-group statistics can blur rather than reveal important information. Worst still, they may even 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 program improvement as an examination of the correlates of failure for the majority.

In sum, the data rarely speak for themselves, but rather 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 program evaluations that are culturally responsive requires people who understand the context in which the data were gathered.

Disseminating and Utilizing 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. 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 (1997) 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.

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.

References

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.

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

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, CA: Jossey-Bass.

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.

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

Patton, M.Q. (1997). *Utilization-Focused Evaluation: The New Century Text*. Thousands Oaks, CA: Sage Publication, Inc.

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. (2001). Understanding and Framing Talent Development School Reform Evaluation Efforts. Talent Development School Reform Evaluation Guide. Washington, DC: Howard University. Center for Research on the Education of Students Placed at Risk (CRESPAR). Unpublished report.

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

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., and Lincoln, Y.S. (eds.). (1994). *Handbook of Qualitative Research.* Thousand Oaks, CA: Sage.

Erlandson, D.A., Harris, E.L., Skipper, B.L., and 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., and Winters, L. (1992). *Tracking Your School's Success: A Guide to Sensible Evaluation.* Newbury Park, CA: Corwin Press.

Hood, S. (forthcoming). Nobody Knows My Name: In Praise of African American Evaluators Who Were Responsive. In *Responsive Evaluation: Roots and Wings*, edited by J. Greene and 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., and 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., and Preissle, J. (eds.). (1992). *The Handbook of Qualitative Research in Education.* San Diego, CA: Academic Press.

Merton, R.K., Fiske, M., and Kendall, P.L. (1990). *The Focused Interview: A Manual of Problems and Procedures*, 2nd Ed. New York: The Free Press.

Miles, M.B., and 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 Art.* 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*. Second Ed. Thousand Oaks, CA: Corwin Press.

Schatzman, L., and 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 Dave 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. [Available online.] http://www.wmich.edu/evalctr/pubs/ops/

Stewart, D.W., and Shamdasani, P.N. (1990). *Focus Groups: Theory and Practice*. Newbury Park, CA: Sage.

The Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards.* Second Ed. Thousand Oaks, CA: Sage Publications.

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.

Yin, R.K. (1989). *Case Study Research: Design and Method.* Newbury Park, CA: Sage.

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 4 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).

Indepth 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 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.

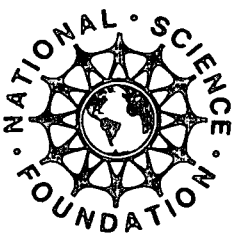
There are at least three avenues that can be pursued:

- 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 in their web site at www.wmich.edu/evalatr. 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.



The National Science Foundation promotes and advances scientific progress in the United States by competitively awarding grants for research and education in the sciences, mathematics and engineering.

To get the latest information about program deadlines, to download copies of NSF publications, and to access abstracts of awards, visit the NSF Web site at:

http://www.nsf.gov

Location:	4201 Wilson Blvd. Arlington, VA 22230
For General Information (NSF Information Center):	(703) 292-5111
TDD (for the hearing-impaired):	(703) 292-5090
To Order Publications or Forms:	
Send an e-mail to:	pubs@nsf.gov
or telephone:	(301) 947-2722
To Locate NSF Employees:	(703) 292-5111

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).

PRESORTED STANDARD
POSTAGE & FEES PAID
National Science Foundation
Permit No. G-69

NSF 02-057

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").