

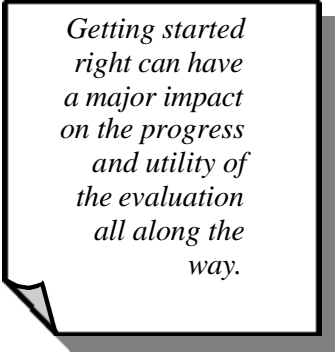
**3. THE EVALUATION PROCESS—
GETTING STARTED**

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.

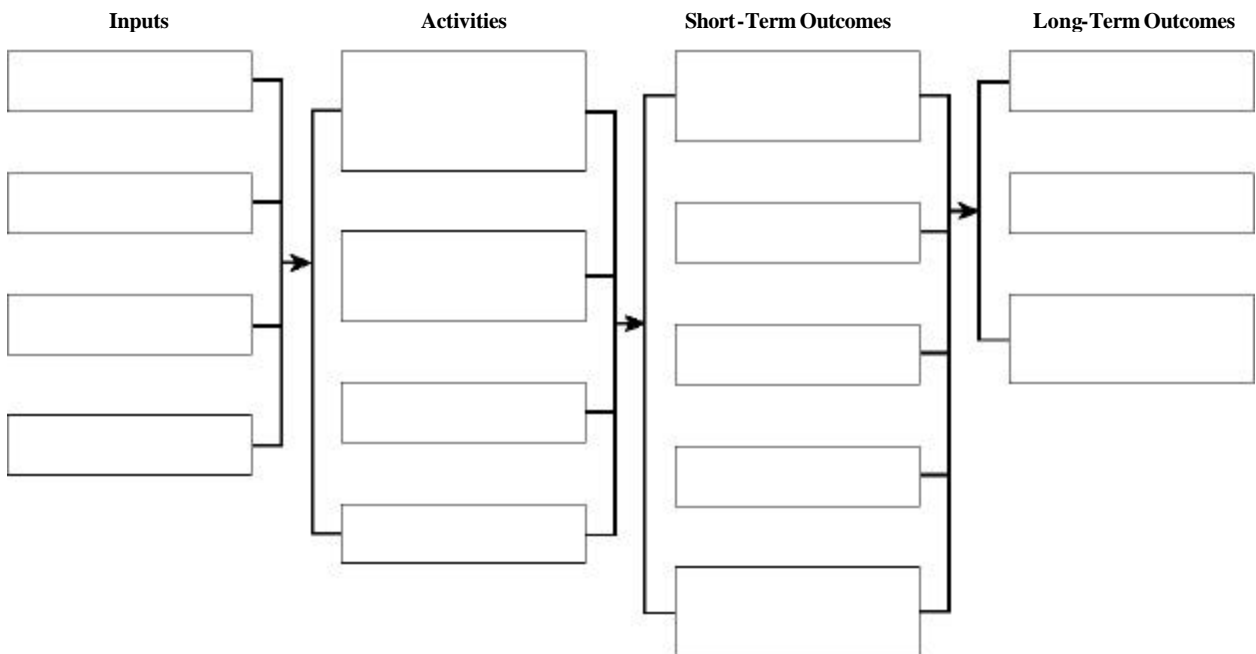
Develop a Conceptual Model of the Project and Identify Key Evaluation Points

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.

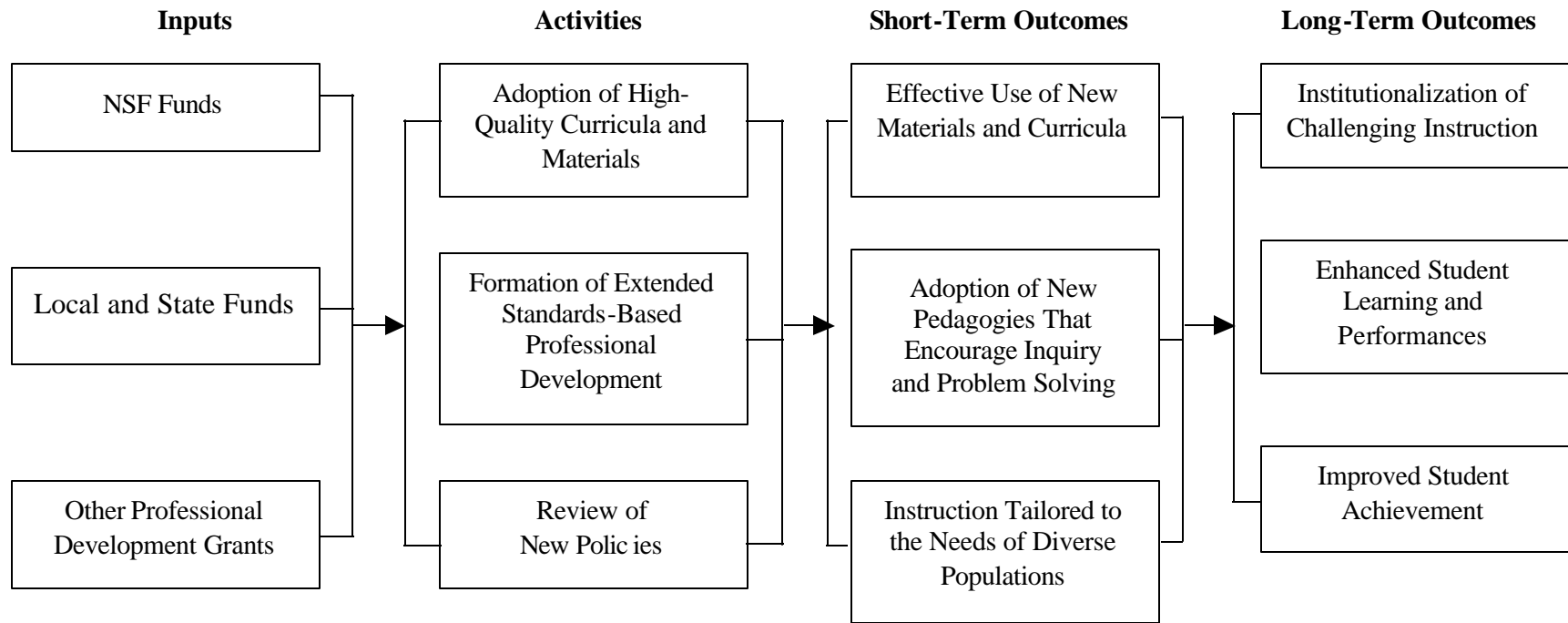
Project inputs are the various funding sources and resource streams that provide support to the project. Activities are the services, materials, and actions that characterize the project's thrusts. Short-term impacts are immediate results of these activities. Long-term outcomes are the broader and more 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 of 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:

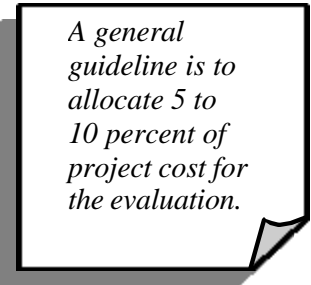
4. Can this objective be broken down further? Break it down to the smallest unit. It must be clear what specifically you hope to see documented or changed.

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



A general guideline is to allocate 5 to 10 percent of project cost for 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.

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.

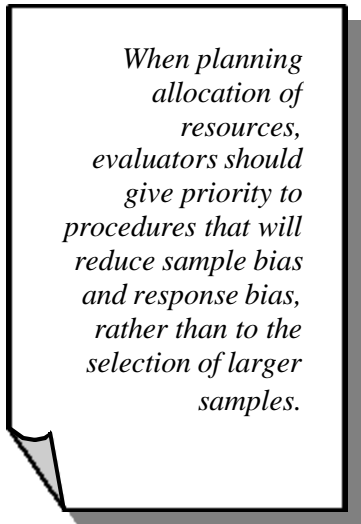
Determining Who Will be Studied and When

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

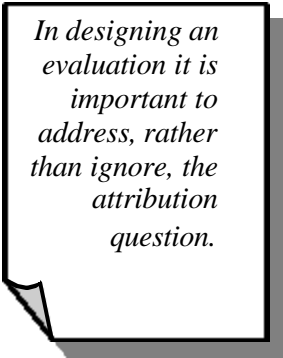
Type	Cause	Remedies
Sampling Error	Using a sample, not the entire population to be studied.	Larger samples—these reduce but do not eliminate sampling error.
Sample Bias	Some of those selected to participate did not do so or provided incomplete information.	Repeated attempts to reach nonrespondents. Prompt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

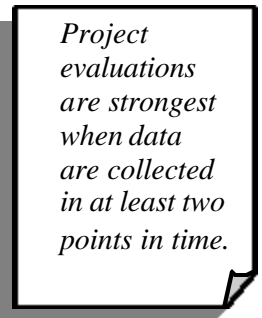
There is no perfect way to solve the problem, but in designing an evaluation it is important to address, rather than ignore, the attribution question. Sometimes this is possible by drawing a comparison group from a waiting list (when one exists) and comparing those who participated with those who self-selected 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

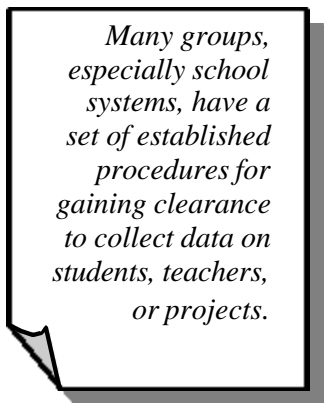
Conducting Data Collection

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.

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



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

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.

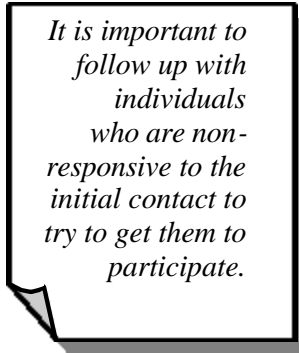
Periodic checks need to be carried out to make sure that well-trained data collectors do not "drift" away from the prescribed procedures over time.

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

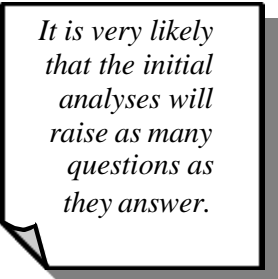
Once the data are collected, they must be analyzed and interpreted. The steps followed in preparing the data for analysis and interpretation differ, depending on the type of data. The interpretation of qualitative data may in some cases be limited to descriptive narratives, but other qualitative data 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.