CHAPTER THREE: DESIGN, DATA COLLECTION AND DATA ANALYSIS

In Chapter Two we outlined the steps in the development and implementation of an evaluation. Another name for that chapter could be "the Soup to Nuts" of evaluation because of its broad-based coverage of issues. In this chapter we focus more closely on selected technical issues, the "Nuts and Bolts" of evaluation, issues that generally fall into the categories of design, data collection and analysis.

In selecting these technical issues, we were guided by two priorities:

> We devoted most attention to topics relevant to quantitative evaluations, because, as emphasized in the introduction, in order to be responsive to executive and congressional decisionmakers, NSF is usually required to furnish outcome information based on quantitative measurement.
>
> We have given the most extensive coverage to topics for which we have located few concise reference materials suitable for NSF/EHR project evaluators. But for all topics, we urge project staff who plan to undertake comprehensive evaluations to make use of the reference materials mentioned in this chapter and in the annotated bibliography.

The chapter is organized into four sections:

- How do you design an evaluation?
- How do you choose a specific data collection technique?
- What are some major concerns when collecting data?
- How do you analyze the data you have collected?

How Do You Design an Evaluation?

Once you have decided the goals for your study and the questions you want to address, it is time to design the study. What does this mean? According to Scriven (1991) design means:

"The process of stipulating the investigatory procedures to be followed in doing a certain evaluation."

Thoughtful analysis, sensitivity, common sense, and creativity are all needed to make sure that the actual evaluation provides information that is useful and credible.

Designing an evaluation is one of those "good news — bad news" stories. The good news is that there are many different ways to develop a good design. The bad news is that there are many ways to develop bad designs. There is no formula or simple algorithm that can be relied upon in moving from questions to an actual study design. Thoughtful analysis, sensitivity, common sense, and creativity are all needed to make sure that the actual evaluation provides information that is useful and credible.

This section examines some issues to consider in developing designs that are both useful and methodologically sound. They are:

- Choosing an approach
- Selecting a sample
- Deciding how many times to measure

Choosing an Approach

Since there are no hard and fast rules about designing the study, how should the evaluator go about choosing the procedures to be followed? This is usually a 2-step process. In step 1, the evaluator makes a judgment about the main purpose of the evaluation, and about the over-all approach which will provide the best framework for this purpose. This judgment will lead to a decision whether the methodology will be essentially qualitative (relying on case studies, observations, and descriptive materials) or whether the method should rely on statistical analyses, or whether a combined approach would be best. Will control or comparison groups be part of the design? If so, how should these groups be selected?

While some evaluation experts feel that qualitative evaluations should not be treated as a technical, scientific process (Guba and Lincoln, 1989) others (for example, Yin, 1989) have adopted design strategies which satisfy rigorous scientific requirements. Conversely, competently executed quantitative studies will have qualitative components. The experienced evaluator will want to see a project in action and

conduct observations and informational interviews before designing instruments for quantitative evaluation; he or she will also provide opportunities for "open-ended" responses and comments during data collection.

There is a useful discussion about choosing the general evaluation approach in Herman, Morris, and Fitz-Gibbon (1987) which concludes with the following observation:

"There is no single correct approach to all evaluation problems. The message is this: some will need a quantitative approach; some will need a qualitative approach; probably most will benefit from a combination of the two."

In all cases, once fundamental design decisions have been made, the design task generally follows the same course in step 2. The evaluator:

- Lists the questions which were raised by stakeholders and classifies them as requiring an Implementation, Progress or Summative Evaluation.
- Identifies procedures which might be used to answer these questions. Some of these procedures probably can be used to answer several questions; clearly, these will have priority.
- Looks at possible alternative methods, taking into account strength of the findings yielded by each approach (quality) as well as practical considerations especially time and cost constraints, staff availability, access to participants, etc.

An important consideration at this point is minimizing interference with project functioning.

An important consideration at this point is minimizing interference with project functioning: making as few demands as possible on project personnel and participants, and avoiding procedures which may be perceived as threatening or critical.

All in all, the evaluator will need to use a great deal of judgment in making choices and adjusting designs, and will seldom be in a position to fully implement text book recommendations. Some of the examples detailed in Chapter Six illustrate this point.

When and How to Sample

It is sometimes assumed that an evaluation must include all of the persons who participate in a project. Thus in teacher enhancement programs, all teachers need to be surveyed or observed; in studies of instructional practices, all students need to be tested; and in studies of reform, all legislators need to be interviewed. This is not the case.

Sampling may be considered or necessary for qualitative and quantitative studies. For example, if a project is carried out in a large number of sites, the evaluator may decide to carry out a qualitative study in only one or a few or them. When planning a survey of project participants, the investigator may decide to sample the participant population, if it is large. Of course, if the project involves few participants, sampling is unnecessary and inappropriate.

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

For qualitative studies, purposeful sampling is often most appropriate. Purposeful sampling means that the evaluator will seek out the case or cases which are most likely to provide maximum information, rather than a "typical" or "representative" case. The goal of the qualitative evaluation is to obtain rich, in-depth information, rather than information from which generalizations about the entire project can be derived. For the latter goal a quantitative evaluation is needed.

For quantitative studies, some form of random sampling is the appropriate method. The easiest way of drawing random samples is to use a list of participants (or teachers, or classrooms, or sites), and select every 2nd or 5th or 10th name, depending on the size of the population and the desired sample size. A stratified sample may be drawn to insure sufficient numbers of rare units (for example, minority members, or schools serving low-income students).

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 which affect the soundness of sample designs. Two other errors—**sample bias** (primarily due to loss of sample units) and **response bias** (responses or observations which do not reflect “true” behavior, characteristics or atti-

tudes)—are much more likely to jeopardize validity of findings. (Sudman, 1976). When planning allocation of resources, evaluators should give priority to procedures which 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 non-response (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 4 describes each type of bias and suggests some simple ways of minimizing them.

Exhibit 4

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 non-respondents. 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 mis-understood, 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.

Determining an adequate sample size sounds threatening, but is not as difficult as it might seem to be at first. Statisticians have computed recommended sample sizes for various populations. (See Fitz-Gibbon and Morris, 1987.) For practical purposes, however, in project evaluations, sample size is primarily determined by available resources, by the planned analyses, and by the need for credibility.

In making sampling decisions, the overriding consideration is that the actual selection must be done by random methods, which usually means selecting every nth case from listings of units (students, instructors, classrooms). Sudman (1976) emphasizes that there are many scientifically sound sampling methods which can be tailored to all budgets:

"In far too many cases, researchers are aware that powerful sampling methods are available, but believe they cannot use them because these methods are too difficult and expensive. Instead incredibly sloppy ad hoc procedures are invented, often with disastrous results."

Deciding How Many Times to Measure

For all types of evaluations (Implementation, Progress, and Summative) the evaluator must decide the frequency of data collection and the method to be used if multiple observations are needed.

For many purposes, it will be sufficient to collect data at one point in time; for others one time data collection may not be adequate. Implementation Evaluations may utilize either multiple or one-time data collections depending on the length of the project and any problems that may be uncovered along the way. For Summative Evaluations, a one-time data collection may be adequate to answer some evaluation questions: How many students enrolled in the project? How many were persisters versus dropouts? What were the most popular project activities? Usually, such data can be obtained from records. But impact measures are almost always measures of change. Has the project resulted in higher test scores? Have teachers adopted different teaching styles? Have students become more interested in considering science-related careers? In each of these cases, at a minimum two observations are needed: baseline (at project initiation) and at a later point, when the project has been operational long

Impact measures are almost always measures of change.

enough for possible change to occur.

Quantitative studies using data collected from the same population at different points in time are called **longitudinal studies**. They often present a dilemma for the evaluator. Conventional wisdom suggests that the correct way to measure change is the "panel method," by which data are obtained from the same individuals (students, teachers, parents, etc.) at different points in time. While longitudinal designs which require interviewing the same students or observing the same teachers at several points in time are best, they are often difficult and expensive to carry out because students move, teachers are re-assigned, and testing programs are changed. Furthermore loss of respondents due to failure to locate or to obtain cooperation from some segment of the original sample is often a major problem. Depending on the nature of the evaluation, it may be possible to obtain good results with successive cross-sectional designs, which means drawing new samples for successive data collections from the treatment population. (See Love, 1991 for a fuller discussion of logistics problems in longitudinal designs.)

There is no hard and fast rule for deciding when changes should or should not be made; in the end technical concerns must be balanced with common sense.

For example, to evaluate the impact of a program of field trips to museums and science centers for 300 high school students, baseline interviews can be conducted with a random sample of 100 students before the project start. Interviewing another random sample of 100 students after the project has been operational for one year is an acceptable technique for measuring project effectiveness, provided that at both times samples were randomly selected to adequately represent the entire group of students involved in the project. In other cases, this may be impossible.

Designs that involve repeated data collection usually require that the data be collected using identical survey instruments at all times. Changing question wording or formats or observation schedules between time 1 and time 2 impairs the validity of the time comparison. At times, evaluators find after the first round of data collection that their instruments would be improved by making some changes, but they do so at the risk of not being able to use altered items for measuring change. Depending on the particular circumstances, it may be difficult to sort out whether a changed response is a treatment effect or the effect of the modified wording. There is no hard and fast rule

for deciding when changes should or should not be made; in the end technical concerns must be balanced with common sense.

How Do You Choose a Specific Data Collection Technique?

In Chapter Two we provided an overview of ways in which evaluators can go about collecting data. As shown in that chapter, there are many different ways to go about answering the same questions. However, the great majority of evaluation designs for projects supported by NSF/EHR rely at least in part on quantitative methods using one or several of the following techniques:

- Surveys based on self-administered questionnaires or interviewer administered instruments
- Focus groups
- Results from tests given to students
- Observations (most often carried out in classrooms)
- Review of records and data bases (not created primarily for the evaluation needs of the project).

The discussion in this section focuses on these techniques. Evaluators who are interested in using techniques not discussed here (for example designs using unobtrusive measures or videotaped observations) will find relevant information in some of the reference books cited in the bibliography.

Surveys

Surveys are a popular tool for project evaluation. They are especially useful for obtaining information about opinions and attitudes of participants or other relevant informants, but they are also useful for the collection of descriptive data, for example personal and background characteristics (race, gender, socio-economic status) of participants. Survey findings usually lend themselves to quantitative analysis; as in opinion polls, the results can be expressed in easily understood percentages or means. As compared to some other data collection methods, (for example in-depth interviews or observations) surveys usually provide wider ranging

but less detailed data and some data may be biased if respondents are not truthful. However, much has been learned in recent years about improving survey quality and coverage and compared to more intensive methods, surveys are relatively inexpensive and easier to analyze using statistical software.

The cheapest surveys are self-administered: a questionnaire is distributed (in person or by mail) to eligible respondents. Relatively short and simple questionnaires lend themselves best to this treatment. The main problem is usually non-response: persons not present when the questionnaire is distributed are often excluded, and mail questionnaires will yield relatively low response rates, unless a great deal of careful preparation and follow-up work is done.

When answers to more numerous and more complex questions are needed, it is best to avoid self-administered questionnaires and to employ interviewers to ask questions either in a face to face situation or over the telephone. Survey researchers often differentiate between questionnaires, where a series of precisely worded questions are asked, and interviews which are usually more open-ended, based on an interview guide or protocol and yield richer and often more interesting data. The trade-off is that interviews take longer, are best done face-to-face, and yield data which are often difficult to analyze. A good compromise is a structured questionnaire which provides some opportunity for open-ended answers and comments.

The choice between telephone and personal interviews depends largely on the nature of the projects being evaluated and the characteristics of respondents. For example, as a rule children should be interviewed in person, as should be respondents who do not speak English, even if the interview is conducted by a bilingual interviewer.

Creating a good questionnaire or interview instrument requires considerable knowledge and skill. Question wording and sequencing are very important in obtaining valid results, as shown by many studies. For a fuller discussion, see Fowler (1993, ch. 6) and Love (1991, ch. 2).

Focus groups

Focus groups have become an increasingly popular information gathering technique. Prior to designing survey instruments, a number of persons from the

population to be surveyed are brought together to discuss, with the help of a leader, the topics which are relevant to the evaluation and should be included in developing questionnaires. Terminology, comprehension, and recall problems will surface, which should be taken into account when questionnaires or interview guides are constructed. This is the main role for focus groups in Summative Evaluations. However, there may be a more substantive role for focus groups in Progress Evaluations, which are more descriptive in nature and often do not rely on statistical analyses. (See Stewart and Shamdasani, 1990 for a full discussion of focus groups.)

The usefulness of focus groups depends heavily on the skills of the moderator, the method of participant selection and last, but not least, the understanding of evaluators that focus groups are essentially an exercise in group dynamics. Their popularity is high because they are a relatively inexpensive and quick information tool, but while they are very helpful in the survey design phase, they are no substitute for systematic evaluation procedures.

Test Scores

Many evaluators and program managers feel that if a project has been funded to improve the academic skills of students so that they are prepared to enter scientific and technical occupations, improvements in test scores are the best indicator of a project's success. Test scores are often considered "hard" and therefore presumably objective data, more valid than other types of measurements such as opinion and attitude data, or grades obtained by students. But these views are not unanimous, since some students and adults are poor test-takers, and because some tests are poorly designed and measure the skills of some groups, especially White males, better than those of women and minorities.

Until recently, most achievement tests were either **norm-referenced** (measuring how a given student performed compared to a previously tested population) or **criterion-referenced** (measuring if a student had mastered specific instructional objectives and thus acquired specific knowledge and skills). Most school systems use these types of tests, and it has frequently been possible for evaluators to use data routinely collected in the schools as the basis for their summative studies.

Because of the many criticisms which have been directed at tests currently in use, there is now a great deal of interest in making radical changes. Experiments with **performance assessment** are under way in many states and communities. Performance tests are designed to measure problem solving behaviors, rather than factual knowledge. Instead of answering true/false or multiple choice formats, students are asked to solve more complex problems, and to explain how they go about arriving at answers and solving these problems. Testing may involve group as well as individual activities, and may appear more like a project than a traditional "test." While many educators and researchers are enthusiastic about these new assessments, it is not likely that valid and inexpensive versions of these tests will be ready for widespread use in the near future.

A good source of information about test vendors and for the use of currently available tests in evaluation is Morris, Fitz-Gibbon and Lindheim (1987). An extensive discussion of performance-based assessment by Linn, Baker, and Dunbar can be found in *Educational Researcher* (Nov. 1991).

Whatever type of test used, there are two critical questions that must be considered before selecting a test and using its results:

- Is there a match between what the test measures and what the project intends to teach? If a science curriculum is oriented toward teaching process skills, does the test measure these skills or more concrete scientific facts?

- Has the program been in place long enough for there to be an impact on test scores? With most projects, there is a start-up period during which the intervention is not fully in place. Looking for test score improvements before a project is fully established can lead to erroneous conclusions.

A final note on testing and test selection. Evaluators may be tempted to develop their own test instruments rather than relying on ones that exist. While this may at times be the best choice, it is not an option to be undertaken lightly. Test development is more than writing down a series of questions, and there are some strict standards formulated by the American Psychological Association that need to be met in developing instruments that will be credible in an evaluation. If at

all possible, use of a reliable and validated, established test is best.

Observations

Surveys and tests can provide good measurements of the opinions, attitudes, skills, and knowledge of individuals; surveys can also provide information about **individual behavior** (how often do you go to your local library? what did you eat for breakfast this morning?), but behavioral information is often inaccurate due to faulty recall or the desire to present oneself in a favorable light. When it comes to measuring **group behavior** (did most children ask questions during the science lesson? did they work cooperatively? at which museum exhibits did the students spend most of their time?) systematic observations are the best method for obtaining good data.

Evaluation experts distinguish between three observation procedures: (1) systematic observations, (2) anecdotal records (semi-structured), and (3) observation by experts (unstructured). For NSF/EHR project evaluations, the first and second are most frequently used, with the second to be used as a planning step for the development of systematic observation instruments.

Procedure one yields quantitative information, which can be analyzed by statistical methods. To carry out such quantifiable observations, subject-specific instruments will need to be created by the evaluator to fit the specific evaluation. A good source of information about observation procedures, including suggestions for instrument development, can be found in Henerson, Morris and Fitz-Gibbon (1987, ch. 9).

The main disadvantage of the observation technique is that behaviors may change when observed. This may be especially true when it comes to teachers and others who feel that the observation is in effect carried out for the purpose of evaluating their performance, rather than the project's general functioning. But behavior changes for other reasons as well, as noted a long time ago when the "Hawthorne effect" was first reported. Techniques have been developed to deal with the biasing effect of the presence of observers: for example, studies have used participant observers, but such techniques can only be used if the study does not call for systematically recording observations as events

occur. Another possible drawback is that perhaps more than any other data collection method, the observation method is heavily dependent on the training and skills of data collectors. This topic is more fully discussed later in this chapter.

Review of Records and Data Bases

Most agencies and funded projects maintain systematic records of some kind about the population they serve and the services they provide, but the extent of available information and their accessibility differ widely. The existence of a comprehensive Management Information System or data base is of enormous help in answering certain evaluation questions which in their absence may require special surveys. For example, simply by looking at personal characteristics of project participants, such as sex, ethnicity, family status etc. evaluators can judge the extent to which the project recruited the target populations described in the project application. As mentioned earlier, detailed project records will greatly facilitate the drawing of samples for various evaluation procedures. Project records can also identify problem situations or events (for example exceptionally high drop-out rates at one site of a multi-site project, or high staff turnover) which might point the evaluator in new directions.

Existing data bases which were originally set up for other purposes can also play a very important role in conducting evaluations. For example, if the project involves students enrolled in public or private institutions which keep comprehensive and/or computerized files, this would greatly facilitate the selection of "matched" control or comparison groups for complex outcome designs. However, gaining access to such information may at times be difficult because of rules designed to protect data confidentiality.

Exhibit 5 summarizes the advantages and drawbacks of the various data collection procedures.

What are Some Major Concerns When Collecting Data?

It is not possible to discuss in one brief chapter the nitty-gritty of all data collection procedures. The reader will want to consult one or more of the texts recommended in the bibliography before attacking any one specific task. Before concluding this chapter, we want to address two issues, however, which affect all data

Exhibit 5

Advantages and Drawbacks of Various Data Collection Procedures		
Procedure	Advantages	Disadvantages
Self-administered questionnaire	Inexpensive. Can be quickly administered if distributed to group. Well suited for simple and short questionnaires.	No control for misunderstood questions, missing data, or untruthful responses. Not suited for exploration of complex issues.
Interviewer administered questionnaires (by telephone)	Relatively inexpensive. Avoids sending staff to unsafe neighborhoods or difficulties gaining access to buildings with security arrangements. Best suited for relatively short and non-sensitive topics.	Proportion of respondents without a private telephone may be high in some populations. As a rule not suitable for children, older people, and non-English speaking persons. Not suitable for lengthy questionnaires and sensitive topics. Respondents may lack privacy.
Interviewer administered questionnaires (in person)	Interviewer controls situation, can probe irrelevant or evasive answers; with good rapport, may obtain useful open-ended comments.	Expensive. May present logistics problems (time, place, privacy, access, safety). Often requires lengthy data collection period unless project employs large interviewer staff.
Open-ended interviews (in person)	Usually yields richest data, details, new insights. Best if in-depth information is wanted.	Same as above (interviewer administered questionnaires); also often difficult to analyze.
Focus groups	Useful to gather ideas, different viewpoints, new insights, improving question design.	Not suited for generalizations about population being studied.
Tests	Provide "hard" data which administrators and funding agencies often prefer; relatively easy to administer; good instruments may be available from vendors.	Available instruments may be unsuitable for treatment population; developing and validating new, project-specific tests may be expensive and time consuming. Objections may be raised because of test unfairness or bias.
Observations	If well executed, best for obtaining data about behavior of individuals and groups.	Usually expensive. Needs well qualified staff. Observation may affect behavior being studied.

collections and deserve special mention here: the selection, training, and supervision of data collectors, and pretesting of evaluation instruments.

Selection, Training and Supervision of Data Collection

Selection

All too often, project administrators, and even evaluators, believe that anybody can be a data collector and typically base the selection on convenience factors: an available research assistant, an instructor or clerk willing to work overtime, college students available for part-time or sporadic work assignments. All of these may be suitable candidates, but it is unlikely that they will be right for all data collection tasks.

Most data collection assignments fall into one of three categories:

- Clerical tasks (abstracting records, compiling data from existing lists or data bases, keeping track of self-administered surveys)
- Personal interviewing (face-to-face or by telephone) and test administration
- Observing and recording observations.

There are some common requirements for the successful completion of all of these tasks: a good understanding of the project, ability and discipline to follow instructions consistently and to give punctilious and detailed attention to all aspects of the data collection. Equally important is lack of bias, and lack of vested interest in the outcome of the evaluation. For this reason, as previously mentioned (Chapter Two) it is usually unwise to use volunteers or regular project staff as data collectors.

Interviewers need additional qualities: a pleasant voice and tactful personal manner and the ability to establish rapport with respondents. For some data collections, it may be advisable to attempt a match between interviewer and respondent (for example with respect to ethnicity, or age.) The need for fluency in a language other than English (usually Spanish) may also be needed; in this case it is important that the interviewer be bi-lingual, with U.S. work experience, so that instructions and expected performance stan-

dards are well understood.

Observers need to be highly skilled and competent professionals. Although they too will need to follow instructions and complete structured schedules, it is often important that they alert the evaluator to unanticipated developments. Depending on the nature of the evaluation, their role in generating information may be crucial: often they are the eyes and the ears of the evaluator. They should also be familiar with the setting in which the observations take place, so that they know what to look for. For example teachers (or former teachers or aides) can make good classroom observers, although they should not be used in schools with which they are or were affiliated.

Training

In all cases, sufficient time must be allocated to training. Training sessions should include performing the actual task (extracting information from a data base, conducting an interview, performing an observation). Training techniques might include role-playing (for interviews) or comparing recorded observations 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.

Supervision

Only constant supervision will ensure quality control of the data collection. The biggest problem is not cheating by interviewers or observers (although this can never be ruled out), but gradual burnout: more transcription errors, more missing data, fewer probes or follow-ups, fewer open-ended comments on observation schedules.

The project evaluator should not wait to review completed work until the end of the data collection, but should do so at least once a week. See Fowler (1991) and Henerson, Morris and Fitz-Gibbon (1987) for further suggestions on interviewer and observer re-

cruitment and training.

Pretest of Instruments

When the evaluator is satisfied with the instruments designed for the evaluation, and before starting any data collection in the field, all instruments should be pre-tested to see if they work well under field conditions. The pre-test also reveals if questions are understood by respondents and if they capture the information sought by the evaluator. Pre-testing is a step that many evaluators "skip" because of time pressures. However, as has been shown many times, they may do so at their own peril. The time taken up front to pre-test instruments can result in enormous savings in time (and misery) later on.

Pre-testing is a step that many evaluators "skip" because of time pressures. However, as has been shown many times, they may do so at their own peril.

The usual procedure consists of using instruments with a small number of cases (for example abstracting data from 10 records, asking 10-20 project participants to fill out questionnaires, conducting interviews with 5 to 10 subjects, or completing half a dozen class room observations). Some of the shortcomings of the instruments will be obvious as the completed forms are reviewed, but most important is a debriefing session with data collectors and in some instances with the respondents themselves, so that they can recommend to the evaluator possible modifications of procedures and instruments. It is especially important to pre-test self-administered instruments, where the respondent cannot ask an interviewer for help in understanding questions. Such pre-tests are best done by bringing together a group of respondents, asking them first to complete the questionnaire, and then leading a discussion about clarity of instructions, and understanding the questions and expected answers.

Data Analysis: Qualitative Data

Analyzing the plethora of data yielded by comprehensive qualitative evaluations is a difficult task, and there are many instances of frequent failure to fully analyze the results of long and costly data collections. While lengthy descriptive case studies are extremely useful in furthering the understanding of social phenomena and the implementation and functioning of innovative projects, they are ill-suited to outcome evaluation studies for program managers and funding agencies. However, more recently, methods have been devised to classify qualitative findings through the use of a special software program (Ethnograph) and di-

verse thematic codes. This approach may enable investigators to analyze qualitative data quantitatively without sacrificing the richness and character of qualitative analysis. Content analysis which can be used for the analysis of unstructured verbal data, is another available technique for dealing quantitatively with qualitative data. Other approaches, including some which also seek to quantify the descriptive elements of case studies, and others which address issues of validation and verification also suggest that the gap between qualitative and quantitative analyses is narrowing. Specific techniques for the analysis of qualitative data can be found in some of the texts referenced at the end of this Chapter.

Data Analysis: Quantitative Data

In Chapter Two, we outlined the major steps required for the analysis of quantitative data:

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

In this chapter, we provide some additional advice on carrying out these steps.

Check the Raw Data and Prepare Data for Analysis

In almost all instances, the evaluator will conduct the data analysis with the help of a computer. Even if the number of cases is small, the volume of data collected and the need for accuracy, together with the availability of PC's and user-friendly software, make it unlikely that evaluators will do without computer assistance.

The process of preparing data for computer analysis involves **data checking**, **data reduction**, and **data cleaning.**

Data checking can be done as a first step by visual inspection of the raw data; this check may turn up responses which are out-of-line, unlikely, inconsistent or suggest that a respondent answered questions mechanically (for example chose always the third response category in a self-administered question-

naire).

Data reduction consists of the following steps:

- Deciding on a file format. (This is usually determined by the software to be used.)
- Designing codes (the categories used to classify the data so that they can be processed by machine) and coding the data. If instruments are "pre-coded," for example if respondents were asked to select an item from a checklist, coding is not necessary. It is needed for "open-ended" answers and comments by respondents and observers.
- Data entry (keying the data onto tapes or disks so that the computer can read them).

Solid data preparation procedures help avoid "GIGO"- garbage in, garbage out.

Many quality control procedures for coding open-ended data and data entry have been devised. They include careful training of coders, frequent checking of their work, and verification of data entry by a second clerk.

Data cleaning consists of a final check on the data file for accuracy, completeness and consistency. At this point, coding and keying errors will be detected. (For a fuller discussion of data preparation procedures, see Fowler, 1991).

If these data preparation procedures have been carefully carried out, chances are good that the data sets will be error-free from a technical standpoint and that the evaluator will have avoided the "GIGO" (garbage in, garbage out) problem which is far from uncommon in analyses based on computer output.

Conduct Initial Analysis Based on the Evaluation Plan

In fact, much can be learned from fairly uncomplicated techniques easily mastered by persons without a strong background in mathematics or statistics.

The evaluator is now ready to start generating information which will answer the evaluation questions. To do so, it is usually necessary to deal with statistical concepts and measurements, a prospect which some evaluators or principal investigators may find terrifying. In fact, much can be learned from fairly uncomplicated techniques easily mastered by persons without a strong background in mathematics or statistics. Many evaluation questions can be answered through the use of descriptive statistical measures, such as frequency distributions (how many cases fall into a

given category), and measures of central tendency (such as the mean or median which refer to statistical measures which seek to locate the "average" or the center of a distribution).

For frequency distributions, the question is most often a matter of presenting such data in the most useful form for project managers and stakeholders. Often the evaluator will look at detailed distributions and then decide on a summary presentation, using tables or graphics. An example is the best way of illustrating these various issues.

Let us assume that a project had recruited 200 high school students to meet with a mentor once a week over a one year period. One of the evaluation questions was: "How long did the original participants remain in the program?" Let us also assume that the data were entered in weeks. If we ask the computer to give us a frequency distribution, we get a long list (if every week at least one participant dropped out, we may end up with 52 entries for 200 cases). Eyeballing this unwieldy table, the evaluator noticed several interesting features: only 50 participants (1/4th of the total) stayed for the entire length of the program; a few people never showed up or stayed only for 1 session. To answer the evaluation question in a meaningful way, the evaluator decided to ask the computer to group data into a shorter table, as follows:

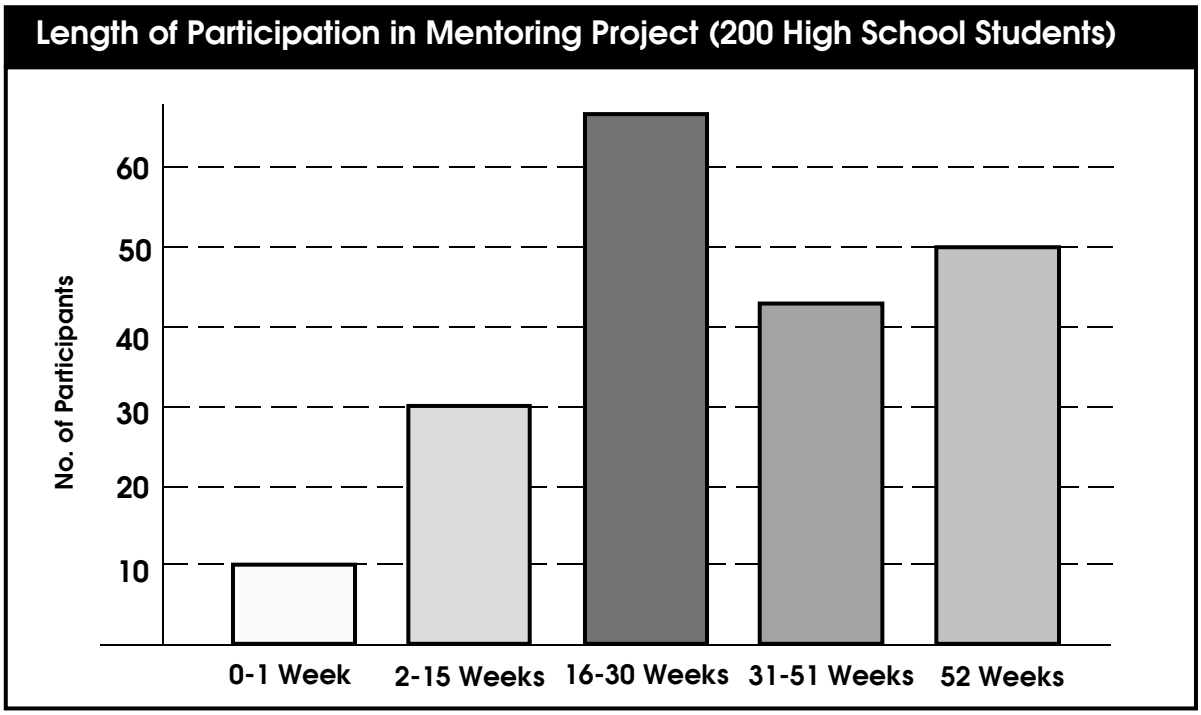
Length of Participation

Time	No. of Participants
1 week or less	10
2-15 weeks	30
16-30 weeks	66
31-51 weeks	44
52 weeks	50

A bar chart might be another way of presenting these data as shown in Exhibit 6.

Let us now assume that the evaluator would like a single figure which would provide some indication of the length of time during which participants remained in the project. There are three measures of central tendency which provide this answer, the mean (or

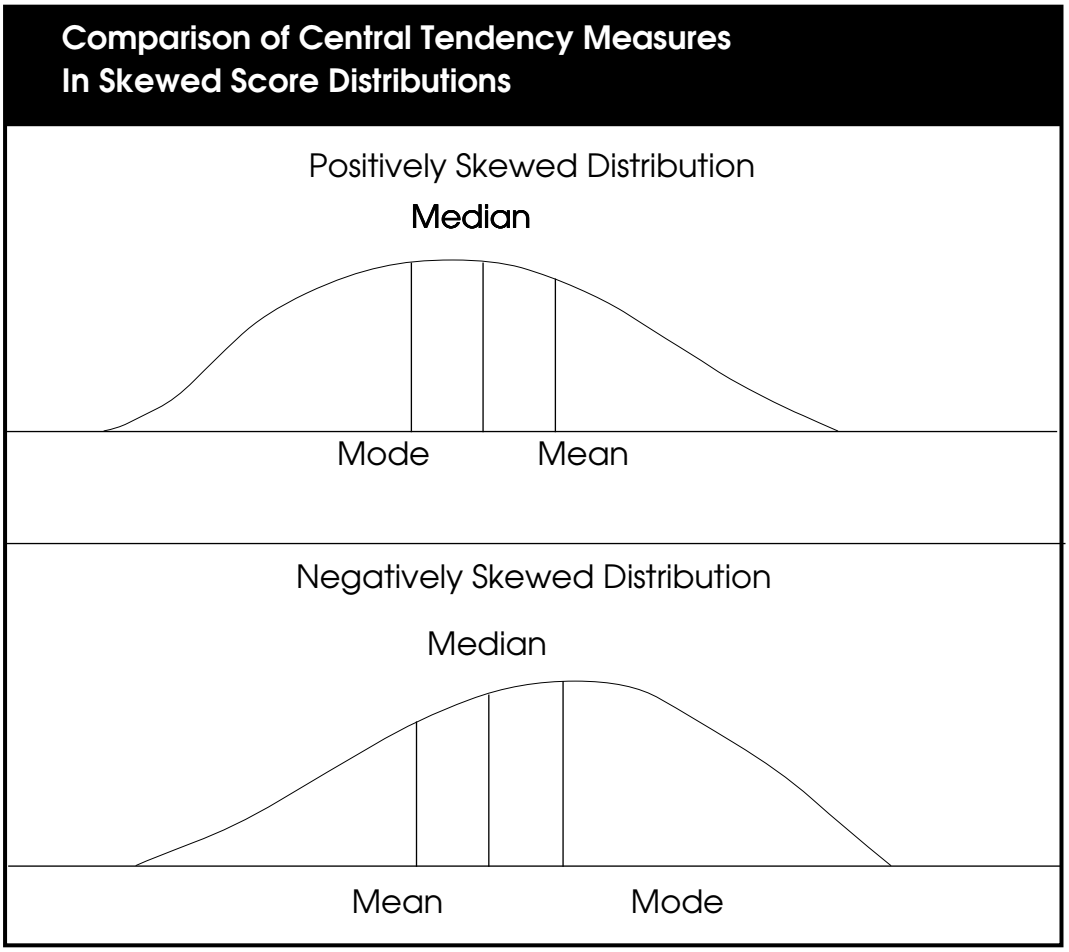
Exhibit 6



arithmetic average), the median (the point at which half the cases fall below and half above), and the mode, which is the category with the largest number of cases. Each of these require that the data meet specific conditions and each has advantages and drawbacks. (See glossary for details.)

In the above example, the only way of computing the mean, median, and mode would be from the raw data, prior to grouping the data as shown in Exhibit 7. However, to simplify the discussion we will just deal with the mean and median (usually the most meaningful measures for evaluation purposes), which can be computed from grouped data. The mean would be slightly above 30 weeks, the median would be slightly above 28 weeks. The mean is higher because of the impact of the last two categories (31-51 weeks and 52 weeks). Both measures are "correct," but they tell slightly different things about the length of time participants remained in the project; the average was 30 weeks, which may be a useful figure for estimating future project costs; half of all participants stayed for 28 weeks or less, which may be a useful figure for deciding how to time retention efforts. Exhibit 7 illustrates differences in the relative position of the median, mean, and mode depending on the nature of

Exhibit 7



the data, such as a positively skewed distribution of test scores (more test scores at the lower end of the distribution) and for a negatively skewed distribution (more scores at the higher end).

In many evaluation studies, the median is the preferred measure of central tendency because for most analyses, it describes the distribution of the data better than the mode or the mean. For a useful discussion of these issues, see Jaeger (1990).

Conduct Additional Analyses Based on the Initial Results

The initial analysis may give the evaluator a good feel for project operations, levels of participation, project activities, and the opinions and attitudes of partici-

pants, staff, and others involved in the project, but it often raises new questions. These may be answered by additional analyses which examine the findings in greater detail. For example, as discussed in some of the earlier examples, it may be of interest to compare more and less experienced teachers' assessment of the effectiveness of new teaching materials, or to compare the opinions of men and women who participated in a mentoring program. Or it might be useful to compare the opinions of women who had female mentors with those of women who had male mentors. These more detailed analyses are often based on **cross-tabulations,** which, unlike frequency distributions, deal with more than one variable. If, in the earlier example about length of participation in mentoring programs, the evaluator wants to compare men and women, the cross-tabulation would look as follows:

One rule of thumb is that a minimum of 20 cases are needed in each subgroup for analysis and for the use of statistical tests to judge the extent to which observed differences are "real" or due to sampling error.

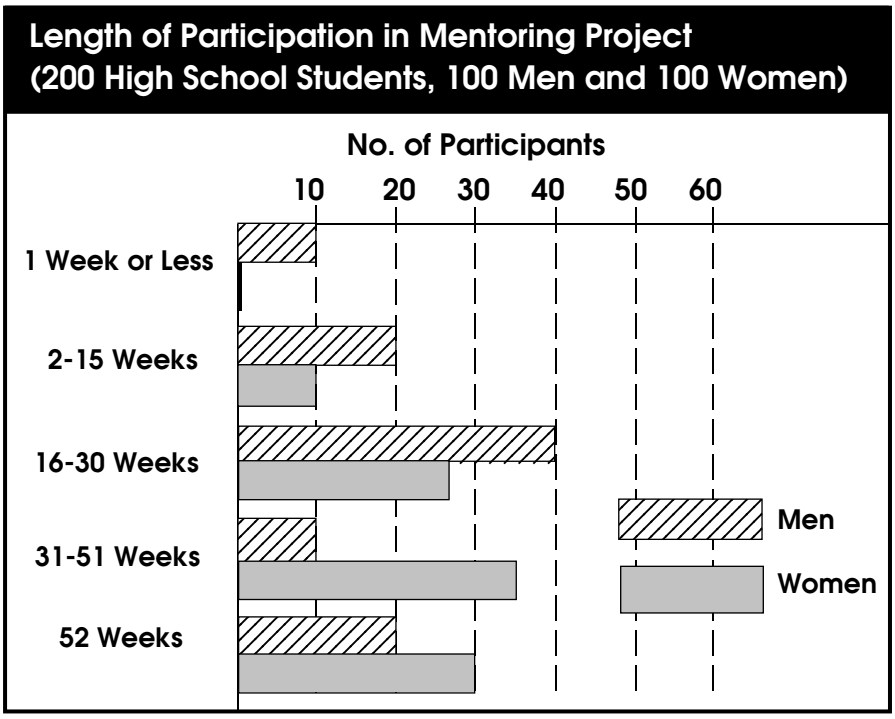
Length of Participation by Sex

	All Students	Men	Women
1 week or less	10	10	0
2-15 weeks	30	20	10
16-30 weeks	66	40	26
31-51 weeks	44	10	34
52 weeks	50	20	30

Exhibit 8, a bar graph, is a better way of showing the same data. Because the table and graph show that on the whole women dropped out later than men, but that most of them also did not complete the entire program, the evaluator may want to re-group the data, for example break down the 31-51 group further to see if most women stayed close to the end of the program.

Cross-tabulations are a convenient technique for examining several variables simultaneously; however, they are often inappropriate because sub-groups become too small. One rule of thumb is that a minimum of 20 cases are needed in each subgroup for analysis and for the use of statistical tests to judge the extent to which observed differences are "real" or due to sampling error. In the above example, it might have been of interest to look further at men and women in different ethnic groups (African American men, African American women, White men and White women) but among the 200 participants there might not have been a sufficient number of African American men or White women to carry out the analysis.

Exhibit 8



There are other techniques for examining differences between groups and testing the findings to see if the observed differences are likely to be "true" ones. To use any one of them, the data must meet specific conditions. Correlation, t-tests, chi-square, and variance analysis are among the most frequently used and have been incorporated in many standard statistical packages. More complex procedures, designed to examine a large number of variables and measure their respective importance, such as factor analysis, regression analysis, and analysis of co-variance are powerful statistical tools, but their use requires a higher level of statistical knowledge. There are special techniques for the analysis of longitudinal (panel) data. Many excellent sources are available for deciding about the appropriateness and usefulness of the various statistical methods (Jaeger, 1990; Fitz-Gibbon and Morris, 1987).

Exploring the data by various statistical procedures in order to detect new relationships and unanticipated findings is perhaps the most exciting and gratifying evaluation task.

Exploring the data by various statistical procedures in order to detect new relationships and unanticipated findings is perhaps the most exciting and gratifying evaluation task. It is often rewarding and useful to keep exploring new leads, but the evaluator must not lose track of time and money constraints and needs to recognize when the point of diminishing returns has been reached.

By following the suggestions made so far in this chapter, the evaluator will be able to answer many questions about the project asked by stakeholders concerned about implementation, progress, and some outcomes. But the question most often asked by funding agencies, planners and policy makers who might want to replicate a project in a new setting is the question: Did the program achieve its objectives? Did it work? What feature(s) of the project were responsible for its success or failure? Outcome evaluation is the evaluator's most difficult task. It is especially difficult for an evaluator who is familiar with the conceptual and statistical pitfalls associated with program evaluation. To quote what is considered by many the classic text in the field of evaluation (Rossi and Freeman, 1993):

"The choice (of designs) always involves trade-offs, there is no single, always-best design that can be used as the 'gold standard'."

Why is outcome evaluation or impact assessment so difficult? The answer is simply that educational projects do not operate in a laboratory setting, where "pure" experiments can yield reliable findings pointing to cause and effect. If two mice from the same litter are fed different vitamins, and one grows faster than the other, it is easy to conclude that vitamin x affected growth more than vitamin y. Some projects will try to measure impact of educational innovations by using this scientific model: observing and measuring outcomes for a treatment group and a matched comparison group. While such designs are best in theory, they are by no means fool-proof: the literature abounds in stories about "contaminated" control groups. For example, there are many stories about teachers whose students were to be controls for an innovative program, and who made special efforts with their students so that their traditional teaching style would yield exceptionally good outcomes. In other cases, students in a control group were subsequently enrolled in another experimental project. But even if the control group is not contaminated, there are innumerable questions about attributing favorable outcomes to a given project. The list of possible impediments is formidable. Most often cited is the fallacy of equating high correlation with causality. If attendance in the mentoring program correlated with higher test scores, was it because the program stimulated the students to study harder and helped them to understand scientific concepts better? Or was it because those who

chose to participate were more interested in science than their peers? Or was it because the school changed its academic curriculum requirements? Besides poor design and measurements, the list of factors which might lead to spurious outcome assessments includes invalid outcome measures as well as competing explanations, such as changes in the environment in which the project operated and Hawthorne effects. On the basis of his many years of experience in evaluation work, Rossi and Freeman (1993) formulated 'The Iron Law of Evaluation Studies':

"The better an evaluation study is technically, the less likely it is to show positive program effects."

There is no formula which can guarantee a flawless and definitive outcome assessment. Together with a command of analytic and statistical methods, the evaluator needs the ability to view the project in its larger context (the real world of real people) in order to make informed judgments about outcomes which can be attributed to project activities. And, at the risk of disappointing stakeholders and funding agencies, the evaluator must stick to his guns if he feels that available data do not enable him to give an unqualified or positive outcome assessment. This issue is further discussed in Chapter Four.

Integrate and Synthesize Findings

When the data analysis has been completed, the final task is to select and integrate tables, graphs and figures which constitute the salient findings and will provide the basis for the final report. Usually the evaluator must deal with several dilemmas:

- How much data must be presented to support a conclusion?
- Should data be included which are interesting or provocative, but do not answer the original evaluation questions?
- What to do about inconsistent or contradictory findings?

Here again, there are no hard and fast rules. Because usually the evaluator will have much more information than can be presented, judicious selection should guide the process. It is usually unnecessary to belabor

a point by showing all the data on which the conclusion is based: just show the strongest indicator. On the other hand, "interesting" data which do not answer one of the original evaluation questions should be shown if they will help stakeholders to understand or seek to address issues of which they may not have been aware. A narrow focus of the evaluation may fulfill contractual or formal obligations, but it deprives the evaluator of the opportunity to demonstrate substantive expertise and the stakeholders of the full benefit of the evaluator's work. Finally, inconsistent or contradictory findings should be carefully examined to make sure that they are not due to data collection or analytic errors. If this is not the case, they should be put on the table, as pointing to issues which may need further thought or examination.

REFERENCES

American Psychological Association, Educational Research Association, and National Council on Measurement in Education (1974). *Standards for Educational and Psychological Tests*. Washington, DC: American Psychological Association.

Fitz-Gibbon, C. T. & Morris, L. L. (1987). *How to Design a Program Evaluation.* Newbury Park, CA: Sage.

Fowler, F. J. (1993). S*urvey Research Methods.* Newbury Park, CA: Sage.

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth Generation Evaluation*. Newbury Park, CA: Sage.

Henerson, M. E., Morris, L. L., & Fitz-Gibbon, C. T. (1987). *How to Measure Attitudes.* Newbury Park, CA: Sage.

Herman, J. L., Morris, L. L., & Fitz-Gibbon, C. T. (1987). *Evaluators Handbook.* Newbury Park, CA: Sage.

Jaeger, R. M. (1990). *Statistics—A Spectator Sport.* Newbury Park, CA: Sage.

Linn, R. L., Baker, E. L., & Dunbar, S. B. "Complex performance-based assessment: expectations and validation criteria." *Educational Researcher*, 20-8, 1991.

Love, A. J. (ed.) (1991). *Evaluation Methods Sourcebook.* Ottawa, Canada: Canadian Evaluation Society.

Morris, L. L., Fitz-Gibbon, C. T., & Lindheim, E. (1987). *How To Measure Performance and Use Tests.* Newbury Park, CA: Sage.

Rossi, P. H. & Freeman, H .E. (1993). *Evaluation—A Systematic Approach* (5th Edition). Newbury Park, CA: Sage.

Scriven, M. (1991). *Evaluation Thesaurus.* Newbury Park, CA: Sage.

Seidel, J. V., Kjolseth, R. & Clark, J. A. (1988). *The Ethnograph.* Littleton, CO: Qualis Research Associates.

Stewart, P. W. & Shamdasani, P. N. (1990). *Focus Groups.* Newbury Park, CA : Sage.

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

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