

**QUANTITATIVELY ORIENTED  
DATA COLLECTION DESIGNS**

Decision making requires relevant data. The problem facing the program and project evaluator is how best (meaning efficiently and effectively) to gather that data. The data must be gathered in a systematic fashion and in such a way as to allow the project impact be seen in as clearly defined form as possible.

It is frequently helpful to formalize a somewhat complex process such as project evaluation into a conceptual paradigm, flow chart, or other schematic. Following is an example of such a schematic.

	<u>Fall</u>		<u>Spring 1995</u>							
	<u>1994</u>									
E (Bulldawg Elementary School)	0 <sub>1</sub>	0 <sub>7</sub>	X	0 <sub>1</sub>	0 <sub>2</sub>	0 <sub>3</sub>	0 <sub>4</sub>	0 <sub>5</sub>	0 <sub>6</sub>	0 <sub>7</sub>
C (Contrast School)	0 <sub>1</sub>	0 <sub>7</sub>	X	0 <sub>1</sub>	0 <sub>2</sub>	0 <sub>3</sub>	0 <sub>4</sub>			0 <sub>7</sub>

Where:

- X = Relevant Instructional Program
- O<sub>1</sub> = Wide Range Achievement Test - Revised (WRAT-R)
- O<sub>2</sub> = Student Survey of Feelings About School
- O<sub>3</sub> = Parent Survey of Perceptions of School Effectiveness
- O<sub>4</sub> = Student Attendance Data
- O<sub>5</sub> = Student Discipline Referral Data
- O<sub>6</sub> = Teacher Sick and Leave Day Data
- O<sub>7</sub> = Elementary Reading Attitude Survey

This schematic represents a typical evaluation design for a school-based innovation project. In this case the innovation was a K--3 continuous progress program in four classrooms. Students at each of the levels were randomly assigned to one of four classrooms. A variety of activities was mounted to maximize the impact of cross-age grouping. Extensive staff development was completed. Curriculum materials were revised or written. Reporting procedures had to be recast in a narrative form. A variety of stakeholders were involved in this evaluation. Their involvement required the

collecting of data from a variety of sources. A comprehensive plan had to be designed. What are the factors that must be considered in selecting or creating a design?

### **FACTORS AFFECTING EVALUATION DESIGN DECISIONS**

The decision to select a specific design, and hence, a specific control or contrast group, involves the weighing of various factors that may impinge upon the project or program due to the specific circumstances surrounding the evaluation and the program. In general, two major influences appear appropriate. First, considerations involving the evaluation design itself must be addressed. Second, practical and political considerations must be assessed.

The usual evaluation design options involve the use of randomization techniques, matching, or the identification of some externally or internally equivalent group. Sometimes fate intervenes to prohibit or at least inhibit the use of any of these three approaches.

What are some of the forces which operate to reduce the likelihood of being able to assign subjects randomly to treatment and control (term used interchangeably with contrast) groups, matched subjects, or to establish equivalent groups?

#### Scope of the Treatment

The scope of the treatment may prevent the use of an optimal evaluation design. The project administrator and evaluator may be faced with a situation where the only politically expedient (or possible) course of action is to assign all students to the treatment group (complete or total coverage). In this case, the only choice with regard to establishing a contrast group is to search for another school or system where matching or establishing a similar group may be possible. Project/program costs and staff preparation may also prove to be additional barriers to selecting a design and a contrast group. For example, where considerable release time for teachers or additional training is required, project administrators may have difficulty in locating (either internally or externally) a sufficient number of participants.

#### Project/Program Purpose

The purpose or purposes of the treatment may affect design considerations. Many programs have as their primary objective the solution of a local problem or a specific set of problems. Although in a general sense one might argue that some projects are developing or testing theory, most agencies require documentation of a need in order for a project to be funded. Should the project administrator be faced with serious deficiencies in student accomplishment or staff performance, the number of potential participants

may greatly increase. The perceived worth (or potential for success) of the treatment may create a condition similar to a bandwagon effect. Systems, classrooms, or schools don't all share the same problem in need of solution.

### Concerns of Parents

Parent support or parent antagonism for projects developed to impact upon students may affect the decision as to which students (if any) are to participate in the project/program. If the potential for solution of a significant problem is high, all parents may want their children to participate. If the perceived potential for harm is high, no parents may allow their children to participate.

### Extent of Treatment

Finally, the extent to which experimental manipulation will occur has a direct bearing on acceptance or participation in the project/program. The project in which only minor changes in routine occur has a greater probability of acceptance than the project in which major changes in routine occurs. In fact, projects calling for major changes in routine may generate sufficient reaction to alter or halt treatment.

From the parent point of view or the teacher point of view, two questions perhaps summarize the dilemma faced by project administrators in the selection of an evaluation design and hence a contrast group: "Who wants to participate as part of the contrast group in a highly successful project?" or "Who wants to participate in the treatment group of a project that is either a flop or is perceived as potentially harmful to the participants?"

## **ELEMENTS OF A DATA COLLECTION DESIGN**

The three major components of a data collection design are included in the foregoing example of the continuous progress program. They are consideration of (1) application of a "treatment" during a particular time frame, (2) the collecting of data from referent groups, and (3) the specification of the data collecting devices. All three of these elements will be dictated by the nature of the problem being investigated and evaluation questions asked. The design simply specifies *what* data are to be gathered from *whom* and *when*. The general nature of the design awaits creation of a more detailed data management plan (see Chapter 8).

### The Use of Contrast Groups

Although there surely are relevant questions surrounding external validity and generalizability and the use of contrast or control groups, the big problem is with internal validity. The basic question is whether we can describe in sufficient detail plausible explanations of the hoped-for differences between groups. So many factors can influence the so-called equivalence of

groups. One need only study the classic "threat list" of Campbell and Stanley (1966) to appreciate that fact (See Table 5-1 later in this chapter). But even randomization is not going to control *all* relevant sources of group contamination. The appeal of randomization technique probably derives from its antecedents in traditional experimental research. Because of that "halo," the technique perhaps has received more accolades than it deserves. The technique can't control all relevant influences. In many instances we in fact want to include those so-called contaminating variables so that their unique interaction becomes part of the "treatment." These "influencing" variables should be free to exert their impact in a naturally occurring environment. The demand characteristics of the evaluation (e.g., student expectations of an improved self-concept) may or may not equate across groups. It is proposed that the term *contrast group* rather than *control group* be used in educational evaluation studies. This term simply refers to an existing or to-be-generated data set against which our "experimental" results are to be contrasted. It is usually the case that in most educational evaluation situations we do not have the luxury of having very extensive control of subjects, or in some cases treatments for that matter. The use of the term *contrast group* would, therefore, be more descriptive of the true state of affairs and in addition tend to remove evaluations from the domain of the traditional experimental paradigm by recasting the nature and focus of the contrast.

Although many evaluation designs are available that do not require the use of contrast or control groups (Cook & Campbell, 1979), most federal and state funding agencies require that such groups be part of the overall data collection and analysis design. The demand for contrast or control groups perhaps reflects an effort on the part of the "money leaders" to force more scientific rigor into the evaluation effort, thereby hopefully generating a more definitive answer to the problem addressed. It also may be perceived that evaluation designs with control groups are more credible and give the appearance of greater validity. Sometimes a norm-referenced external data base might be used, but in contemporary evaluation practice there is a definite affinity for classical experimental designs.

Horan (1980) has suggested that historically we have considered control groups to have received "everything but" the experimental treatment. It may be truer to say that in far too many instances "anything but" might be a better descriptor. But it is agreed that evaluation involves some kind of comparison. That benchmark might come from data generated from a contrast group in the design or be derived from an extant source (e.g., test manual statistics.) *A major criterion for almost any good evaluation design is the use of independent contrast data.* The comparison may be to some like-type group without the prescribed treatment or it may be some extant data base such as a set of national or state norms. The concern is to design our evaluations so



that, within practical limits, rival hypotheses may be ruled less plausible. The realistic evaluator is less likely than the "brass instruments" researcher or the experimental design obsessive to be concerned with causal inference.

Once a project administrator and evaluator have assessed the circumstances and have determined the evaluation design and the contrast group, other questions bearing upon the effective use of the contrast group must be addressed. One such question involves payoffs for both the participants and the decision makers in the contrast group: "What will we get from this experience?" If the contrast group receives no benefits in regard to program, professional development, or other kinds of rewards, a reluctance to participate can probably be anticipated. Project administrators must also decide what kinds of information will be presented to the participants and decision makers of the contrast group. In general, it would appear that the contrast group should receive all of the feedback from all measurements taken in the same time and manner as the treatment group. Finally, the issue of competition must be considered. Where the contrast group resides outside of the school or district, old rivalries may stir up a competitive attitude. Beware of the John Henry effect where the control outperforms the experimental. For example, a project involving students in two high schools (one constituting the control) where athletic competition has been keen in the immediate past may be affected by transfer of the competitiveness to the objectives of the project. One obvious method for avoiding this situation is to select a control school where there is no history of keen competition with the treatment school. Other possibilities to avoid the influence of competition include selective statistical analysis (e.g., ANCOVA), use of project data, and other information.

### Categories of Designs

Three general categories of designs will be considered here: (1) experimental, (2) quasi-experimental, and (3) nonexperimental. These three classes of design differ in the degree of control over the treatment that they allow. We are attempting to isolate and measure the impact of our program or project. We want, in essence, to hold constant as many as possible and feasible extraneous factors and influences that might "contaminate" our results. Pedhazur and Schmelkin (1991) note that there are four major methods of exerting control in the design of studies.

First and foremost is *randomization*. Randomization as used here refers to the process of selecting or assigning whatever the sampling unit and ultimate analysis unit is (e.g., individual student, teacher, classroom, school) to a condition (e.g., competing treatments, a treatment and a control) so that each unit will have an equally likely chance of being in each of the conditions.

Chance will determine placement. Tables of random numbers or computer programs can be used effectively to accomplish randomization. Although less efficient and perhaps not useful with extremely large data sets, such manual methods as flipping coins, rolling dice, or drawing numbers from a hat can be used. An approximation of random selection can be accomplished by randomly entering a list of names or identification numbers and then taking every  $n$ th name as needed. The intent is to "equate" groups so that everyone begins on the same footing and that any potential factors that might influence the outcome measures, independent of the treatment, are controlled or at least confounded (i.e., don't have a systematic effect). There are some evaluators who don't believe in randomization. They say that rare events can and do happen. Yes, they do, but only rarely!

Doing project evaluations in the real world usually does not allow for the luxury of employing complete randomization. In the foregoing example of the continuous progress program, although the students were randomly assigned to classrooms, the contrast school was not randomly selected along with the experimental school. There are a limited number of *statistical controls* available that will help us make adjustments for the lack of equivalence between the two schools. A very powerful technique is analysis of covariance (ANCOVA). This procedure allows post intervention scores or an outcome measure to be adjusted for initial differences between an experimental and contrast (control) group. The adjusting variable (covariate) is usually a premeasure equivalent or similar to the post measures, but any variable(s) thought or known to be correlated (statistically and conceptually) with the dependent or outcome measure could be used. One of the important corollary benefits of using ANCOVA is greater power and precision in the analyses. Partial correlation is another technique useful in holding constant control variables. We might, for example, be interested in the relation between scores on the Graduate Record Examination and graduate school grade point average, holding constant or controlling for age correlationally. Another approach would be simply to run the correlations separately for different age groups.

The actual selection of the treatment(s) in the programs and projects represents another method of control. The choice of intervention in intensity and duration can be controlled or *manipulated* thereby allowing for an assessment of its impact.

Finally the independent and extraneous variables can be controlled by *including* or *excluding* them from influence. Variables that might be hypothesized to be related to the outcome measures or treatment can be controlled by selection. Variables such as sex, age, race, or socioeconomic class can be controlled by limiting a study to a particular group (e.g., females) or the evaluation could be conducted using separate but intact groups (e.g.,

all females versus all males). The variable of gender would thereby be held constant. Using this procedure may, of course, limit the generalizability of the results. The technique is particularly useful when the anticipated influencing variable is categorical.

When one thinks of the myriad of variables that can influence the results of any evaluation, randomization must be considered as an effective control mechanism, particularly for large samples. Failing that, do the best you can with the other methods, but always be cautious in interpreting your results.

### THE VALIDITY OF DATA COLLECTION DESIGNS

The literature of classical experimental research is replete with caveats to the investigator about all the factors that can mess up the results. Campbell and Stanley (1963), Cook and Campbell (1979), and more recently Campbell (1986) have helped several generations of investigators understand threats to design validity. The original set aggregated by Campbell and Stanley (1963) included internal and external validity that reflected on the control of the treatment and generalizability of the results, respectively. Cook and Campbell (1979) added statistical conclusion and construct validity to the list. These related to the inferences from statistical tests and the treatment-outcome measure match, respectively. Campbell (1986) has changed slightly the focus and interpretation of internal and external validity. Internal validity has been renamed Local Molar Causal Validity. Translated, that means that there is greater emphasis on controlling the extraneous complex interacting factors that influence implementation of the project at the local level. There is also greater concern now for the theoretical relationship between the treatment and the outcome measures. The external validity concept has been recast as Proximal Similarity. The renaming of this concept is an attempt to capture the uniqueness of treatment-site interaction. Selecting a representative sample for the evaluation may not be as important as describing the conceptual and actual interaction of treatments, measures, populations, settings, and times. Exportability will then be to those environments where there is greatest similarity. Documentation of the experimental and contrast environments is, therefore, an essential element in the design process.

Because of the familiarity of the evaluation community with the original labels of internal and external validity, we will continue to use them here.

What are the threats to design validity and how can they be controlled?

Table 5-1 contains a summary of nine significant factors that can distort (either positive or negative) the evaluation of a program or project.

Any one of the influences described in Table 5-1 can be further confounded by interactions with any of the other influences. Interactions with *selection* in particular can be particularly detrimental to design validity.

**TABLE 5-1 Summary of Threats to Internal Validity of Data Collection Designs**

<u>Category</u>	<u>Description</u>	<u>Example</u>
History	Events related to outcome of study occur during implementation.	Local outbreak of AIDS occurs during conduct of AIDS awareness program in high school.
Maturation	Naturally occurring uncontrolled changes in subjects that are related to outcome.	Elementary school physical education program shows increased skill development although it could be simply due to aging.
Testing	Repeated data collection may result in increased scores. Operation of practice or memory.	Short duration of attitude toward drug program in middle school requires pre- and post-measurements to be gathered only weeks apart.
Statistical Regression	A real phenomenon where post-treatment scores of those at extremes move toward "average."	At-risk preschoolers selected because of low scores on screening tests show significant gains after one year of intervention.

<u>Category</u>	<u>Description</u>	<u>Example</u>
Instrumentation	Change in instrumentation over course of study. Changes in calibration or scoring accuracy.	Lack of comparability in two forms of high school chemistry test used to assess impact of lab-oriented curriculum.
Mortality	Attrition of study subjects occur at higher rate for experimental group or contrast group.	During three-year project aimed at enhancing English skills of Hispanics finds that more of the contrast than the experimental group has left the area.
Selection	Differential or self-selection biases groups.	Volunteer schools in the contrast group tended to come from low socioeconomic areas in a study of the impact of a self-esteem building program.
Diffusion/imitation of Treatments	A competing treatment to that voluntarily adopted by the experimental group is adopted by the contrast group.	The contrast teachers also attend the staff development sessions on cooperative learning methods meant for the experimental group.

<u>Category</u>	<u>Description</u>	<u>Example</u>
Compensatory Rivalry/resentful Demoralization	Differential effort as a result of real or perceived differential treatment.	The contrast school, not having received a computer lab, tries harder to do a better job teaching elementary math.

Validity is the control of the treatment effect. Take, for example, the possible interaction of selection and history. Due to a defect in the selection or assignment process, for example, more upper socioeconomic students got into the experimental program. A measure of progress in a language arts program might be enhanced artificially, as another example, because of greater availability of academic support mechanisms; books, computers, and so on. It is in fact these interactions, particularly of the treatment with (1) selection, (2) history, and (3) the setting in which the evaluation takes place, that contribute to decreased generalizability (for external validity) of the results. Lack of control, one of the extraneous factors, or a high degree of uniqueness of the group(s) or subject(s) contribute to the difficulty in replicating the results.

One can see how important it is to monitor the treatment as it is being implemented. Lack of fidelity in application of the innovative treatment can totally confound the results. One should in fact *evaluate* the implementation of the treatment.

#### Concern for Unintended Effects

Another design issue relates to the problem of unintended program effects. It is here where perhaps using a goal-free approach makes a great deal of sense. An evaluator might say, "Don't bias my data gathering by telling me what you expect; let me see what happened for myself." Wolf (1984) has likened the search for unintended outcomes to looking for a black cat in a dark room on a moonless night. It is almost that difficult but also important. We can be both happily surprised or depressed with unintended outcomes. The present chapter began with a description of a continuous progress nongraded program at the elementary school level. Among the unexpected negative outcomes of this project was the finding that kindergarten discipline referrals were greatly increased relative to previous years. This was attributed to the fact that teachers in the experimental classrooms had to deal with four age groups (5,6,7, and 8). The first six months of a kindergartner's school life can be traumatic. The continuous-progress teacher had to deal with *all*

children and perhaps had less opportunity to deal with the kindergartners' special problems. On the positive side was the finding that students saw themselves as special and experienced a concomitant increase in self-esteem. How did these unexpected effects become apparent? Primarily through observation, *ex post facto* examination of discipline referral data, and focus group interviews with students (see Chapter 6). Observation of the in-progress program is a particularly useful method of data gathering related to implementation.

What implications do these factors have for the actual design of an evaluation? In the following section nine data collection designs will be presented and the various advantages and disadvantages discussed.

### DATA COLLECTION DESIGNS

It was noted in the previous section that "control" was the key to a good design. If one cannot or does not want to randomize, then other methods might be employed. In the language of experimental psychology the application of randomization procedures should result, for example, in the creation of equivalent groups. One will receive our treatment (or experience the innovative program or project) and the other will act as a reference point, benchmark or comparison group against which data from the experimental group can be contrasted. Rarely can we randomize and get a *control* group after the fact. As was noted earlier, the alternative term *contrast* group is suggested. Every effort will be made to make the groups comparable. Perhaps data from records and files could be used. In the illustrative data collection design presented at the beginning of this chapter, it was found that the school means on the state criterion referenced tests in reading and mathematics were within three points of each other and that the percent of students on free or reduced lunch was 63% for Bulldawg Elementary and 58% for the contrast school. The judgment of assumed comparability was made.

In doing educational evaluation there never exists a setting where a "no treatment" condition exists. Everybody gets something, some more or less than others. Is it better to give than to receive? There is always a "traditional treatment" going on. It may not be systematic or continuous, but it exists. One of the tasks that an evaluator must complete then is to describe *both* the experimental and contrast treatment. Comparing the applications of these two treatments is what it's all about.

The traditional design symbology will be employed here: X = a treatment and O = an observation, measurement, or data collection event. A convention will be used here, however, where observations with the same subscript will mean the same or equivalent measurement e.g., parallel test forms, no matter when they are taken. Consider:

$$O_1 \times O_1$$

In this case the same measurement was used on a pre-treatment/post-treatment basis.

One final introductory comment is that we are here specifying differences between experimental and quasi-experimental designs on the basis of a failure to apply randomization procedures in the quasi-experimental situation. In addition, quasi-experimental designs are differentiated from nonexperimental (sometimes referred to as pre-experimental) designs because the later do not reflect any randomization or manipulation of a treatment variable.

### Experimental Designs

Our first experimental design is the ever popular and usually effective *Pre-test--Post-test Contrast Group Design*.

$$\begin{array}{l} \text{R Group 1} \quad O_1 \quad X \quad O_1 \\ \text{R Group 2} \quad O_1 \quad \quad O_1 \end{array}$$

Randomization has been accomplished (R). Multiple observations either pre or post could be made if so desired, and they usually are. Because of randomization the major threats to internal validity have been controlled. We will obviously only complete our analyses on subjects (students) who were present for the entire study. Using randomization controls for regression and selection effects and the pre-test allows for examination of the effect of mortality. In addition, presence of a contrast group controls for history, testing, and instrumentation. Finally, the combination of randomization and the presence of a contrast group control for maturation. Interpretation of results is reasonably straightforward. The design could obviously be expanded to include several different treatment groups. Extending the basic two-group design to multiple groups and measurements might yield a configuration such as the following:

$$\begin{array}{l} \text{R Group 1} \quad O_1 \quad X_A \quad O_1 \quad O_2 \quad O_1 \\ \text{R Group 2} \quad O_1 \quad X_B \quad O_1 \quad O_2 \quad O_1 \\ \text{R Group 3} \quad O_1 \quad \quad O_1 \quad O_2 \quad O_1 \end{array}$$

Such an extension would allow us to examine competing treatments (and perhaps conduct cost analyses; see Chapter 8) and do follow-up studies (the third  $O_1$ ). In addition, the introduction of an observation ( $O_2$ ) which the



evaluator felt was a measure of a relevant outcome could be accomplished after the treatment, thereby avoiding any chance of testing X treatment interaction or sensitization (reactivity.)

A possible weakness of this design is the presence of potential interaction between pre-test and treatment. Subjects might be "sensitized" to the treatment simply having taken a pre-test. An attitude toward drugs inventory might cause students to think about their knowledge and feelings regarding this topic even before an educational program was completed. In that sense the pre-test becomes part of the treatment. Students might seek more information on their own or converse at length with their peers about their reaction to the problem.

A useful design to control for the treatment of pre-testing interaction is the *Post-Test--Only Contrast Group Design*, represented as follows:

R	X	$O_1$
R		$O_1$

There is no pretest included in the design. Again, multiple treatments and observations of a number of different outcome (dependent) variables could be gathered. The design does not allow for assessing the effect of "mortality" because it lacks a pre-test. If the size of the sample is large and the duration of the study is relatively short, then mortality will probably not be a factor. Remember that one of the meanings of control rests on an evaluator's ability to assess the effect on uncontrollable extraneous influences even if they can't actually manipulate the variable.

If it is important for the evaluator in fact to gauge the amount of gain or change over the duration of the study on a measure that poses a potential pre-test--interaction threat, then perhaps the "mother" of all designs, the *Solomon Four Group Design*, could be used.

It is represented as follows:

R Group 1	$O_1$	X	$O_1$
R Group 2	$O_1$		$O_1$
R Group 3		X	$O_1$
R Group 4			$O_1$

Groups are constituted by random assignment to one of four separate units. Two groups are pre-tested, and one of them receives the treatment. They are

both post-tested. What you have, of course, is our old friend the Pre-test--Post-test Contrast Group Design. One of the remaining two groups is post-tested and one of these receives the treatment. What we have here is another old friend, the Post-test--Only Contrast Group Design. Contrasting Groups 1 and 3 in the above diagram will allow us to assess the impact of the pre-test-treatment interaction (and mortality) if it was generated. The Solomon Four Group Design enjoys all the advantages of the Pre-test--Post-test Contrast Group Design and the Post-test--Only Contrast Group Design. Drawbacks of the design are that it (1) requires a lot of units (e.g., students, classrooms), and (2) is not very practical in public school settings.

#### Controlling for Reactivity with Retrospective Pre-testing

Often it is not possible to identify an acceptable control or contrast group. In addition, when sensitive treatments are involved (e.g., attitudes toward drug use) pretesting may generate a so-called pre-test effect which reacts with the dependent measure. To accommodate these difficulties, Campbell and Stanley (1966) have suggested the use of *retrospective pre-testing*. Such a procedure allows the treated group to act as its own control, a particularly useful approach when self-report dependent measures are involved (Howard *et al.*, 1979). An allied problem is the phenomenon of "response-shift bias." Assume for a moment that you are going to be a participant in a workshop on problem-solving skills. The pre-test contains an item like the following: "I am a good problem solver." You strongly agree with the statement and so respond. After getting into the workshop you find that you really aren't a very effective problem solver. At the end of the intervention you are confident about the skills you have developed and again, but for a different reason, strongly agree with the statement, "I am a good problem solver." Obviously, a "no-difference" conclusion would be reached when evaluating the workshop. One method of "finding" some relevant contrast data involves, as the title suggests, actually gathering *ex post facto* pre-test data. One could, for example, have our workshop participant fill out an end-of-workshop questionnaire, providing a summative evaluation of its effects and values. The participant would then be asked to respond to the questionnaire as s/he would have if s/he had taken it prior to the experience. (Often it is not physically or operationally possible to gather pre-test data. A few years ago the author was involved in evaluating the State of Georgia Governor's Honors Program (GHP) for the academically and artistically talented. This enrichment experience for rising high school juniors and seniors lasted for eight weeks in a college campus setting. The lack of availability of anything remotely resembling a contrast group was evident. Several post-experience measures were gathered which basically served the same purpose as pre-testing. Students, for example, were asked to contrast

their summer experience with the regular school treatments. Illustrative are the following questions:

Which holds the student more responsible for work?

In which do students try out their ideas more?

Which provides greater opportunity for close contact with teachers?

Possible answers were: GHP; Regular school; No difference.

Retrospective testing has been used primarily with affective measures, but some researchers have used the technique successfully in the cognitive domain (Rippey, Geller, & King, 1978). In our GHP evaluation students were asked, for example, to make retrospective judgments about the extent to which the program contributed to their mastery of selected instructional objectives.

So much for experimental design. What do I do if I cannot randomly assign units to conditions?

#### Quasi-Experimental Designs

It was noted that the public school project and program evaluator usually do not have the opportunity to apply randomization procedures to evaluation studies. A very frequently employed design that approximates within certain parameters the *Pre-test--Post-test Contrast Group Design* is the *Non-equivalent Contrast Group Design*. The only difference between this design and the former is that randomization procedures have **not** been used.

Group 1	$O_1$	X	$O_1$
	.....		
Group 2	$O_1$		$O_1$

The dotted line between the groups indicates lack of randomization. Two or more groups might be employed and, again, multiple measures possibly used. The lack of randomization allows for the influence of sources of invalidity not present with the pre-test--post-test contrast group design--namely, regression, and possible interaction between selection and variables such as maturation, history, and testing. Since the groups are not equivalent, a frequently used statistical procedure is analysis of covariance. In an effort to help ensure the closest similarity between the experimental and contrast group(s), matching procedures are sometimes used. One method of "constructing" a contrast group has been suggested by Payne and Brown (1982).

### Constructing Matched Groups

Although it is not held in the highest regard by all quantitative methodologists, the use of matching procedures can provide meaningful contrast data useful in assessing evaluation data. An argument against matching is that for every variable on which individuals or groups are matched there may be many others of equal or greater importance. Despite this potential shortcoming, matching can be a valuable design technique. The following method, the *Aggregate Rank Similarity Method*, was first described by Brown (1980) and involves the matching of an experimental classroom, school, or school system against a population of possible contrast groups. A more elaborate system has been described by Sherwood, Morris, & Sherwood (1975). One distinct advantage of the matching procedure is that the evaluator has control of the variables, and in many cases data on the more important ones are readily available. Matching can take place either within or outside a district. Let's look at an example to illustrate the procedure.

An evaluator is interested in identifying a school system to use as a contrast system while evaluating a new K--12 science curriculum that involves integrating both career education and environmental concerns. Before a list of matching variables is generated, a sample of potential contrast groups is identified which are geographically contiguous to the "experimental" system. A list of independent (matching) variables is then assembled. Criteria for inclusion in the list could be (1) justification for the relevance of the matching variable found in the research literature, and (2) availability of on-site data. Table 5-2 contains three such variables. The raw data for each variable for each potential contrast system are subtracted from the values of the data for the experimental system, and the *absolute* values entered. Next these absolute values are ranked from smallest to largest for each variable. The ranks are summed across the variables and the system with the smallest aggregate sum

**TABLE 5-2 Illustration of Aggregate Rank Similarity Method for Matching Systems**

SYSTEM	MATCHING VARIABLES									
	<u>% of Pupils on Free Lunch</u>			<u>Average Daily Attendance</u>			<u>Per Pupil Expenditure</u>			
	<u>Raw Data</u>	<u>d'</u>	<u>Rank</u>	<u>Raw Data</u>	<u>d</u>	<u>Rank</u>	<u>Raw Data</u>	<u>d</u>	<u>Rank</u>	
EXPERIMENTAL	76			6284			4395			
Contrast 1	62	14	2	4387	1897	3	5031	636	2	7
Contrast 2	60	16	3	5934	350	1	3977	418	1	5
Contrast 3	98	22	4	3847	2437	4	3828	567	4	12
Contrast 4	74	02	1	5166	1118	2	5127	732	3	6

\*Note: Absolute deviation of contrast system data from experimental data.

is selected as the contrast system. In the case of the data in Table 5-2 it would be System 2. A great variety of independent (matching) variables might be identified; for example, scores on standardized tests, pupil/teacher ratios, data from observation instruments, or credentialing/certification data. Selection is limited only by the creativity of the evaluator and project administrator. In addition, the selected independent matching variables might be differentially weighted.

The *Interrupted Time Series Design* is another valuable quasi-experimental design. It is particularly useful when large amounts of comparable archival data are available. Impact is assessed by examining the change in measurements after the introduction of an innovation or treatment. The basic design such as this

$$0_1 \quad 0_1 \quad 0_1 \quad X \quad 0_1 \quad 0_1 \quad 0_1$$

can be augmented with an equivalent or quasi-equivalent contrast group. When used with a contrast group it is referred to as a multiple time series design. Used in the multiple form it might look like this:

Group 1	$0_1$	$0_1$	$0_1$	$X$	$0_1$	$0_1$	$0_1$
Group 2	$0_1$	$0_1$	$0_1$		$0_1$	$0_1$	$0_1$

The interrupted time series design controls for the history and instrumentation threats to internal validity, which were not controlled in the single form of the design. This design can be used effectively when the collection of repeated measures (e.g., test scores, attendance data, disciplinary referrals) is an ongoing and naturally occurring activity. It is particularly useful when the entire population must receive the treatment (i.e., complete coverage with the intervention).

A final quasi-experimental design is the *Institutional Cycle Design*. This design (a variation on the counterbalanced design) is again useful when all subjects must receive the treatment. A school, for example, wants all elementary students to experience a new environmental AIDS awareness unit. We could take the entire elementary student body (probably using the classroom as the unit) and assign half of them to the first implementation of the environmental unit. They would experience the unit (X) and then be post-tested. The second group would take the environmental post-test as a pre-test; then they would experience the unit and be post-tested.

The data collection design would look as follows:

Group 1	X	0 <sub>1</sub>	(Y)	0 <sub>2</sub>			
Group 2		0 <sub>1</sub>	X	0 <sub>1</sub>	0 <sub>2</sub>	(Y)	0 <sub>2</sub>

We have two measures of program impact. Group 1 Post versus Group 2 Pre, and Group 2 Post versus Pre. The design could be jazzed up by adding another treatment (Y) to each group. The design is an interesting combination of cross-sectional and longitudinal approaches. The design does, however, suffer from a failure to control for three problems: maturation, selection, and possible multiple treatment interactions.

#### Data Analysis from Nonequivalent Contrast Group Designs

It is not possible to address this enormous analysis topic here with the space limitations at hand. Suffice it to say that there are many technical issues involved in evaluating data from nonequivalent contrast group designs, although it may be as Lord (1967) has said: "With the data usually available for such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups." There do exist, however, a number of useful analysis procedures that can be applied to data generated from quasi-experimental designs (Reichardt, 1979). Such techniques as analysis of covariance, value-added analysis, regression-discontinuity, and selection-modeling are illustrative valuable data analysis procedures. In addition, the reader is alerted to the volumes noted in the Suggested Readings section at the end of this chapter for current information about analysis procedures appropriate for most of the frequently used data collection designs: Freed (1991), Keppel (1991), and Pedhazur and Schmelkin (1991).

#### An Illustrative Nonequivalent Contrast Group Evaluation Study

It was a dark, rainy, thundering, overcast, dreary, stormy day when the aspiring evaluator was called into the department head's office. As chairperson of the reading education department, she had been contacted by a major instructional materials developer and solicited to serve as a field-testing site for a new set of six computer-assisted instruction modules aimed at teaching the teaching of reading techniques to prospective teachers. There were eight sections of the relevant course available next quarter. Enrollment averaged about 20--25 students per class. The dean and department head requested that she take on this project (for no extra compensation, as is usually the case) citing the benefits of visibility for the department and college,

possible publications, and presentations at professional meetings. The evaluator met with the developer of the computer-assisted instruction modules, department head, and dean, and evolved the following modest evaluation questions:

1. Will the new CAI reading education modules result in increased knowledge about techniques for the teaching of reading?
2. Will the new CAI reading education modules result in greater learning than that resulting from using the current traditional techniques?
3. Will the new CAI reading education modules result in positive attitudes toward computer-assisted instruction about the teaching of reading?

There are virtually an infinite number of possible designs that could have been created to answer the three evaluation questions. What follows is one of many possible approaches. Among considerations involved in creating this design and the ultimate report were the:

1. Desire to utilize already existing classes.
2. Need to gather contrast data from CAI and non-CAI groups.
3. Relatively short period of time in which to conduct the study.
4. Lack of funds to develop elaborate instrumentation.
5. Inability to control assignment of students to classes at registration due to scheduling needs.

### Design

Since the evaluation questions required both *absolute* information (questions 1 and 3) and *relative data* (question 2) a nonequivalent contrast group design was used. Four classes were designated as CAI and four non-CAI. Naturally occurring enrollment determined which class a student attended.

In addition, periodic monitoring of how well instructors were using the new materials was undertaken to maintain fidelity of implementation. An in-service program had been held by the materials developer on the use of the CAI materials with the four CAI instructors.

The design could be symbolized as follows:

CAI Classes	$O_1$	$O_2$	$X_c$	$O_1$	$O_2$	$O_3$	$O_4$
Non-CAI Classes	$O_1$	$O_2$	$X_n$	$O_1$	$O_2$	$O_3$	
Where	$O_1$	=	Pre and post measures of knowledge of teaching of reading techniques				
	$O_2$	=	Pre and post measures of attitude toward computer assisted instruction				
	$O_3$	=	Immediate post-test on individual modular units				
	$O_4$	=	Interviews with selected CAI students				
	$X$	=	Treatment (Interventions, Instruction, Project Program) C = CAI, N = Non-CAI (Traditional)				

It should be noted that although we were unable to randomly assign students to classes (and thereby treatments), the only way of looking at growth is by getting some kind of change data over time, and therefore we had to use pre and post measures and a contrast group.

#### Data-Gathering Instruments

The Reading Techniques Knowledge Inventory (RTKI) was a 75-item, 5-alternative, multiple-choice test designed specifically for the project. It had a pilot-test-calculated Kuder-Richardson Formula 20 internal consistency reliability of .73. The inventory contained content questions as well as problem application exercises. The items were scored right or wrong (and yes or no).

The attitude scale employed was a 17-item inventory using a 5-point Likert rating scale (Strongly Agree. . . Strongly Disagree). Following is a sample item:

Computers are one of the best ways to teach.

This instrument, Attitude Toward Computerized Instruction (ATCI), had a reported test-retest reliability of .87 over a four-week period and had been shown in other studies to be moderately related to actual classroom performance.

Each of the six curricular modules had a 25-item summative test covering only the material of that unit. They were administered to individual students when they had concluded each unit. The items were scored right or wrong. The CAI modules dealt with the same content in the same sequence as in the traditional classes.



### Data Collection and Storage

Modern optical scanning technology allows for the collection and processing of very large amounts of data. The data were scanned directly into a micro computer. Item booklets with standard optical scan sheets were sent to each instructor. Data were gathered from all participants using the same directions.

### Data Sources

Students in the two groups (CAI and non-CAI) were instructed with their relevant respective materials for one 11-week quarter. Part of the first week was given over to organizational and orientation matters and pre-testing, and part of the last week was devoted to post-instruction data collection. The CAI materials involved six modules spread over a 10-week period. Each CAI lesson required about two and a half hours of working time. Students could work at their own pace. The CAI material was supportive of ongoing instruction and represented approximately 50% of the total instructional time.

### Decision Rule/Standard Setting

The use of a decision rule is a way of incorporating the concept of standard/criteria setting (see Chapter 3). Several approaches could be taken. One could specify the percentages of specific instructional objectives that need to be mastered, either by individuals or groups, as a means of rendering a decision about effectiveness, or an overall mean score difference could have been specified. Another approach, the one taken in the present evaluation, is to use a statistical model to help make the judgments about effectiveness. A combination of descriptive data together with inferential statistical methods was used.

### Data Analysis

The data collection design suggested the following kinds of analyses.

In order to answer Evaluation Question 1, the pre and post scores were contrasted for each of the module unit tests and for the RTKI. A correlated t-test was used (a test of differences between means for a single group of students).

Evaluation Question 2 required application of analysis of covariance on the pre--post modular and RTKI scores across the two groups (CAI vs. non-CAI classes). Use of this particular procedure allowed potential differences between the two groups before the new program was introduced to be adjusted or "equalized."

The final evaluation question dealing with differences in attitude between groups was assessed through application of a t-test of differences between correlated means for the same students.

Additional analyses could have been specified concerning differences between the effectiveness of the program for males as opposed to females, or the effect of amount of familiarity with computers on achievement and attitude.

### Results

Table 5-3 contains a summary of the means and standard deviations of the RTKI scores. The class was used as the unit of analysis. This was done because of the potential unique interaction between instructor and student. The percent score was obtained by dividing the means by the number of dichotomous items (75).

**TABLE 5-3 Summary of Pre-Test, Post-Test, and Mean Score Differences for Reading Techniques Knowledge Inventory CAI and Non-CAI Groups**

	Pre-Test				Post-Test			Mean Gain
	<u>n</u>	<u>Mean</u>	<u>%</u>	<u>SD</u>	<u>Mean</u>	<u>%</u>	<u>SD</u>	
CAI	4	51.32	68	8.88	64.73	86	7.33	13.41*
Non-CAI	4	49.98	67	10.78	54.34	72	6.07	4.36
Differences		1.34			10.38*			9.05*

\*This difference significantly greater than zero,  $p < .05$ . (Some analysts prefer simply to report the actual p-values).

The following interpretations appear justified from these data:

1. There were no initial (pre-test) differences between the CAI and non-CAI classes.
2. There was a meaningful knowledge gain for the CAI group but not the non-CAI group.
3. There were meaningful differences between the two groups at the end of the quarter (post-test).
4. The gains were significantly greater for the CAI than the non-CAI classes.

It appears that the CAI materials had a positive influence on knowledge acquisition. Whatever the reason--ability to self-pace, opportunity for review, or periodic within module testing--the CAI delivery system brought about enhanced learning. Note also that the variability of scores became less at the end of instruction. Such a phenomenon might be interpreted as reflecting positively on the reliability of program implementation--in other words, students were more alike at the end of the instructional experience than they were at the beginning.

Table 5-4 contains a summary of the total scores and subscores (by objective) for Module One. These are presented here as an illustration of the kinds of analyses carried out for each module. The subscores are tied to five general objectives, each of which was measured by five items. The data are presented simply to illustrate the kind of information that can be gathered. Such data can be used formatively to improve the instructional materials.

Following are the instructional objectives for the first module:

1. Associate different instructional practices with three conceptual frameworks (models) of the reading process.
2. Apply phonics generalization to decode nonsense words.
3. Identify the purposes of Language Experience Activity.
4. Apply syllabication rules.
5. Recognize the characteristics of a particular Language Experience Activity.

It can be seen that the module is generally working in the hoped-for way, with the total scores being higher for the CAI group. Two of the subparts of the module are perhaps in need of attention--namely, subparts 2 and 4. Referring back to the objectives we can see that objectives 2 and 4 tend to be a bit more technical and complex than the others. This is a situation where the formative approach to evaluation will help us *improve* the curricular materials. Performance on the items related to objective 2 suggest that the content, materials, or instructional approach are not effective based on the low level of achievement. The comment also could be made about objective 4 where students are not performing well against an absolute criterion. In any event, improvements are needed.

**TABLE 5-4 Summary of Means and Standard Deviations of Total and Subscores on Module One Test**

Objective	Group	Mean	SD
<u>One</u>	CAI	3.14	1.11
	Non-CAI	2.38	.93
<u>Two</u>	CAI	2.67	1.24
	Non-CAI	4.01	1.86
<u>Three</u>	CAI	4.37	.88
	Non-CAI	3.72	1.32
<u>Four</u>	CAI	2.87	1.77
	Non-CAI	1.65	.56
<u>Five</u>	CAI	3.99	.74
	Non-CAI	1.44	.66
<u>Total</u>	CAI	17.01	5.36
	Non-CAI	13.40	4.98

The attitude data are summarized in Table 5-5. The rating scale means for this 17-item instrument can range from 17 to 85. The attitude data again favor the CAI approach. There is significant gain over time as well as a differential gain across groups. The absolute level end-of-treatment attitude is pretty positive at the conclusion of the quarter.

In summary, what do our data tell us about our 10-week instructional intervention? It appears that it works cognitively and that attitudes also have been positively influenced. The data also suggest that improvements can yet be made in the CAI materials and the instructional approaches used in selected modules. Other changes could be suggested by item analysis of the data from all instruments used in the study.

Education is concerned with the realization and utilization of human resources. Measurement and evaluation can significantly aid in their realization and utilization by providing reliable and valid data on where we have been, where we are, where are we headed, and how much we have accomplished.

**TABLE 5-5 Summary of Pre-Test and Post-Test Means and Standard Deviations for Scores on the Attitude Toward the Computer-Assisted Instruction Instrument**

	Pre-Test			Post-Test		Mean Gain
	<u>n</u>	<u>Mean</u>	<u>SD</u>	<u>Mean</u>	<u>SD</u>	
CAI	4	49.34	8.31	62.30	7.70	12.96*
Non-CAI	4	48.01	7.11	53.41	6.69	5.40
Differences		1.33		8.89*		8.56*

\*Significantly different from zero,  $p < .05$ .

#### Nonexperimental Designs

This general class of designs is sometimes referred to as pre-experimental designs. They do not allow for the manipulation of the treatment or randomization. The designs are nevertheless helpful at the formative level in field testing materials or procedures. The first nonexperimental design is the *One-Shot Case Study*:

$$X \quad 0_1$$

The evaluator simply describes outcomes apparently resulting from the application of a fixed treatment. The term *design* is used advisedly here as none of the major threats to validity are controlled. A well-done case study, if implemented by an accomplished qualitative evaluator (see Chapter 6), can nevertheless yield invaluable information about the impact of a program or project. Among the potential benefits of the case study method is the generation of hypotheses or questions to be answered in future studies. The evaluator must be extremely careful about drawing inferences from this approach. Its framework is exploratory at best. See books by Merriam (1988) and Yin (1984) for comprehensive treatments of the case study method.

A second nonexperimental design is the *One Group Pre-test--Post-test Design*:

$$0_1 \quad X \quad 0_1$$

Its use involves pre- and post-testing a single group which has received a particular treatment. Rossi and Freeman (1989) refer to this design as a *reflexive* control design since the treatment group serves as its own control. The single group time-series design can also be called a reflexive design.

Uncontrolled factors in this design include history, maturation, testing, instrumentation, and selection interactions with a variety of other factors. If the time between pre- and post-observations is lengthy, these threats have the potential for significant impact. The inability to assess potential pre-test-treatment reactivity is a serious drawback of this design. As with the case study method, use of this design might be helpful during the early stages of product or project development. Again, be warned about drawing meaningful inferences from this design since by its very nature it results in rival hypotheses in most cases being more tenable than usual.

The *Static Group Comparison* is a slightly improved nonexperimental design. A case-study-like design is supplemented with some contrast data.

Group 1	X	$O_1$
	-----	
Contrast Data		$O_1$

We might compare, for example, the general equivalency diploma (GED) scores of a group of Hispanic adults who had been working on a computer tutorial program preparing them to take the high school equivalency exam with a mean score for recent high school graduates. The normative test group provides the contrast group data. Another set of contrast or reference data could be derived from what Rossi and Freeman (1989) call a *shadow* control. Judgments from experts or program participants are gathered to serve as benchmarks for interpreting impact data. We might ask a group of experts to fill out a teacher evaluation form in a manner that would describe an "effective multicultural teacher." This profile could be used as a reference point for evaluating the impact on individual teachers of a staff development program. Still another kind of contrast group is sometimes referred to as a *generic* control. Extant data bases, such as the normative GED scores referred to earlier, can be used as comparisons against the outcomes of a particular intervention. State, local, or national indices are available through a variety of public and private agencies and publications.

In concluding this chapter, several observations need emphasis:

- All effective evaluation designs require the collection of contrast, benchmark, or comparison data.
- Evaluative designs evolve. The relationship between problem/question and method of seeking an answer is inseparable.

- Considerations of cost, nature of questions, nature of administrative and political constraints, and receptivity of decision maker(s) will influence evaluability (the viability of even doing the evaluation at all).
- The key to an effective evaluation design is the isolation and measurement of the impact of the treatment, intervention, or innovation.
- Don't let sampling procedures or statistical methods dictate the evaluation design.
- Generalizability is nice, but internal validity is essential.
- If approached in a systematic way, using accepted guidelines, intelligence, and common sense, any evaluation design can yield valuable information.
- Although technical aspects of the evaluation method are important, all will fail if the problem is not properly conceptualized.
- Anticipating the things that can go wrong and planning for their rectification or control is half the battle in the war for truth.

### COGITATIONS

1. What are the major considerations that differentiate true experimental, quasi-experimental, and nonexperimental designs?
2. What design factors should be most important to the (a) evaluator, and (b) project director?
3. What designs would be best to use when the entire population of targets must be covered or served?
4. What are the advantages and disadvantages of using reflexive, constructed (matched), generic, or shadow contrast groups?
5. What is "experimental" about an experimental design? What is "quasi" in a quasi-experimental design? What is "pre" about a pre- or nonexperimental design?
6. Are the factors that influence design creation the same as those that effect design selection?
7. Are some threats to internal validity more important than others? Why?
8. What are the advantages and disadvantages of retrospective pre-testing as an approach to controlling testing by treatment interaction?

9. Under what conditions can nonexperimental studies be valuable?
10. Under what conditions would unintended effects be acceptable?

### SUGGESTED READINGS

- Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally. The modern classic treatise.
- Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation design and analysis issues for field settings*. Chicago: Rand McNally. What do you do when you can't randomize?
- Freed, M.N. (Ed.) (1991). *Handbook of statistical procedures and their computer applications to education and the behavioral sciences*. New York: American Council on Education/Macmillan. The introduction to the major microcomputer software statistics packages (e.g., SAS, SPSS-X, SYSTAT, MINITAB) is particularly informative.
- Keppel, G. (1991). *Design and analysis (A researcher's handbook)*. (3rd ed.) Englewood Cliffs, NJ: Prentice Hall. Very comprehensive, but not for the fainthearted.
- Mohr, L.B. (1992). *Impact analysis for program evaluation*. Newbury Park, CA: Sage. Twelve designs in all their glory. Food for thought and action.
- Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design and analysis (An integrated approach)*. Hillsdale, NJ: Lawrence Erlbaum. Eight hundred nineteen pages and 24 chapters of all one could possibly want to know, but the authors really communicate.
- Popham, W.J. (1993). *Educational evaluation*. (3rd ed.) Boston: Allyn and Bacon. Chapter 10 contains a realistic overview of a variety of usable designs and hints on their implementation.
- Taylor Fitz-Gibbon, C., & Morris, L.L. (1987). *How to design a program evaluation*. Newbury Park, CA: Sage. Easy to read yet comprehensive decision-trees help readers find their way in the forest of designs.
- Trochim, W.M.K. (Ed.) (1986). *Advances in quasi-experimental design and analysis* (New Directions for Program Evaluation, No. 31). San Francisco: Jossey-Bass. Six important papers representing current conceptions and issues.



## QUALITATIVE AND ETHNOGRAPHIC EVALUATION

Mary Jo McGee-Brown  
University of Georgia

*We do a lot of looking: we look through lenses, telescopes, television tubes. . . Our looking is perfected every day--but we see less and less.*

F. Franck (1973, p. 3)

Rist (1980), in his article "Blitzkrieg Ethnography: On the Transformation of a Method into a Movement," expresses a number of concerns about the decline in quality of the conceptualization, process, and products of ethnographic research in education simply because the approach was "in vogue." Rist claims that "The term 'ethnographer' is now being used to describe researchers who neither studied nor were trained in the method." Rist raises other concerns such as researchers conducting hit-and-run research rather than designing and conducting longitudinal studies; researchers having simple description as the end in itself rather than exploration of the underlying cultural framework and deep meanings of participants; multiple-researcher multisite research focusing on breadth rather than single-ethnographer, single-site designs focusing on in-depth understanding; and entrance into a research site with preformulated research problems and concepts resulting in a predetermined approach to data collection and analysis rather than allowing them to emerge from extensive time and interaction with participants at the site. Rist predicted that as the number of untrained researchers and evaluators employing this method grows, the rationale for using the qualitative approach will be undercut, the resulting reports will be of poor quality, and disenchantment with the qualitative approach will be inevitable.

We are at a similar place in evaluation today. It is common for funders to want to know "what is happening out there" when they provide support for programs. That question requires a qualitative approach to data collection. It is generally assumed that untrained persons cannot conduct quantitative data collection, statistical analysis, and data interpretation. On the other hand, many have the misconception that "anyone can do qualitative research and evaluation" because the methods include observation and interviewing. While it is admirable that evaluators are looking to expand their tools, we are moving toward a disaster if qualitative evaluators are not systematically trained in the philosophical and theoretical assumptions underlying the approach, field strategies, design issues, data collection methods, and data analysis and

interpretation approaches. The author knows of one situation in a state where project evaluators are trained in a two-day workshop, an hour of which focuses on qualitative evaluation. The coordinator then feels that all workshop participants (most of whom have a quantitative background) participating in that workshop were qualified to design and implement qualitative components in their evaluations. Nothing could be further from the truth!

### A RATIONALE FOR INTERPRETIVE INQUIRY

The underlying assumption of qualitative evaluation is that the perspectives and actions of all participants or stakeholders in a program are important. There are four primary reasons for selecting a qualitative approach in evaluation. They are to:

- *Discover* the meanings that the innovation, program or project has for persons across levels within at the site(s).
- *Observe* the effects of the innovation or change on behavior, actions, and interactions for all persons at the site(s).
- *Document* the process in the natural setting without manipulating any variables.
- *Assess* cultural changes that are a direct or indirect result of the program as well as determine the effects of the larger cultural context on changes associated with the program.

Qualitative inquiry is an umbrella term that includes many different research designs. The term *interpretive inquiry* is preferred (see Figure 6-1), because understandings from all qualitative methods of data collection by nature include multiple levels of interpretation. As Erickson (1986) notes, interpretive inquiry is more inclusive than qualitative, and it avoids the connotation that quantification is not used in interpretive research.

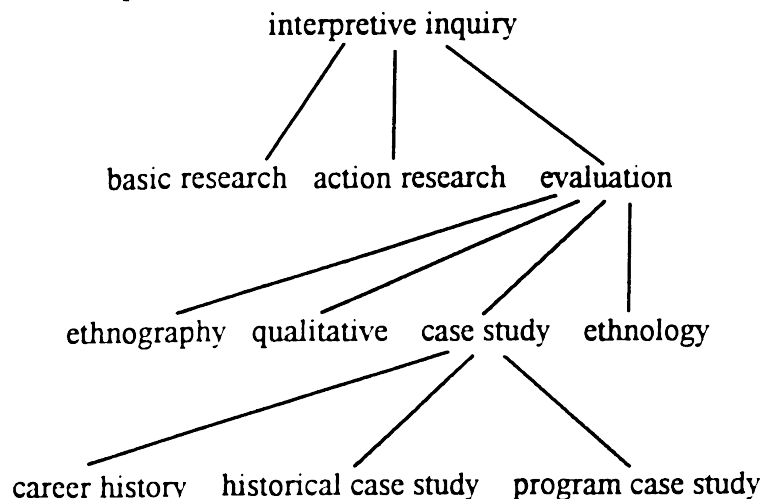


Figure 6-1 Interpretive Evaluation Designs

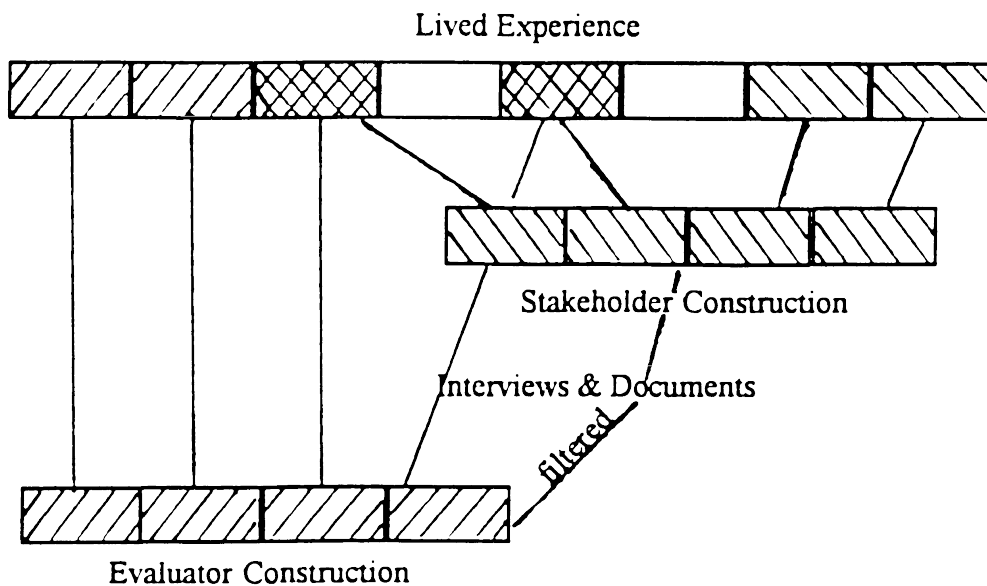
The key features of each of the interpretive evaluation designs are that (1) they are longitudinal, (2) there is a fieldwork component in which the evaluator collects data in context at the site(s) while experiencing the program first-hand with participants, (3) data are generally narrative in form (but numeric data are also collected), (4) there is a focus on finding participants' meanings for the program in that context, and data are analyzed inductively. The interpretive evaluator is seeking participants' reactions to and evaluations of a program and reasons for those evaluative perspectives. Qualitative evaluators use the interpretive inquiry mode.

There are multiple levels of interpretation of social reality. The first level is the interpretation of the lived experience by all those who live it. The second level of interpretation is a result of the pieces of the lived experience that remain in the memory of a participant. The third level of interpretation is what is selected out of those memories to share with the evaluator at any given point of data collection. Each participants' interpretations are filtered through different lenses which are constructed of all previous knowledge, experiences, and beliefs of that individual as a result of being a part of a particular culture at some given point in time. The most important caution given to qualitative or ethnographic evaluator is, therefore, that "There may be a correspondence between a life as lived, a life as experienced, and a life as told, but the anthropologist should never assume the correspondence nor fail to make the distinction" (Bruner, 1984).

Figure 6-2 illustrates the social "reality" that is constructed by individuals who share experiences.

Geertz (1973) suggests that all ethnography involves only second and third order interpretation, claiming that only "natives" can make first order interpretations of their culture. While the "natives" in a context, or program participants, can make first order interpretations, the participant observer functioning as evaluator tries to approximate that level of interpretation by assuming a role in the social organization so that s/he experiences the program first-hand along with participants from the site. The evaluator brings her/his own subjectivity to the situation, and from that subjectivity constructs meaning about the program. By interacting with participants, they construct meaning together. The interpretive evaluator builds a second order interpretation through careful systematic observations, ongoing conversational and informal structured interviews, and ongoing participant documentation based on a commitment to the importance of revealing what is "actually happening" in the program and a total trust of the evaluator with their data. Longitudinal designs where the evaluator interacts closely with participants

allows the evaluator to understand the circumstances in which participants would tell, share, or reveal something other than what they understand to be the "real" lived experience in the program. When this "deception for survival" (Brown, 1991) occurs in the evaluation context, understanding *why* this is happening is critical to accurate interpretation of *what* is happening. Usually an understanding of the larger cultural context and power differences provides the insights for interpreting the difference between lived experiences and told experiences.



**Figure 6-2 Individual Constructions of Social Reality**

Note in Figure 6-2 that only parts of the lived experience are remembered by the stakeholders and used in their retrospective construction of what the experience was like. The same process occurs with the evaluator who functions as participant observer. The evaluator's construction is further impacted by the information collected through interviews with stakeholders and documents about a program. Interview data are filtered further through the types of questions the evaluator poses, the selection process by the stakeholder of what to share with the evaluator, and the evaluator selection of data to include as examples in the final report. Note also that parts of the lived experience have been totally forgotten by all participants.

An ethnographic evaluator functions as a participant observer at the evaluation site and simply "lives" with the participants as much as possible to both experience and interpret the innovation or program and systematically observe the effects of it in context as a basis of a cultural analysis. A qualitative evaluator uses the tools (participant observation, interviewing, and document analysis) of ethnography but does not necessarily stay on site continuously nor conduct a cultural analysis of the program (Fetterman, 1984).

## QUALITATIVE DESIGN ISSUES

You might ask yourself, "How do I begin to create a good qualitative evaluation design?" The immediate answer to this question is to begin by being open to finding out *how* a program is affecting all persons involved and *why* it affects different levels of persons differently. These two broad goals for the evaluation can guide all data collection and analysis as you revisit them periodically. While you begin holistically, the outcomes of your inquiry will dictate the narrowing of the focus based on participants' perspectives of what is important rather than you as the evaluator defining the focus prior to entering the site. Structured flexibility and openness are key to a good qualitative evaluation design.

### Unit of Analysis Problems

In designing qualitative evaluations, the first thing to consider is identification of the unit(s) of analysis. Patton (1990, p. 168) asserts that "The key issue in selecting and making decisions about the appropriate unit of analysis is to decide what it is you want to be able to say something about at the end of the study." Determine what information policy makers need about a program and from whom they need it. Following this, a careful examination of the social organization will lead to identification of the sampling process within the selected site(s) which would address the unit of analysis.

As can be seen from the example in Exhibit 6-1, many evaluation decisions are based on the determination of the unit of analysis. The sampling process, the sample, data collection methods, and the number of evaluators needed on the team are all affected by the identification of the unit of analysis. The evaluator, not the stakeholders, will determine the unit of analysis, but it cannot be done without gaining a clear understanding of what the stakeholders want from the evaluation.

It is important to note that the evaluation can have more than one unit of analysis. The evaluator may want to be able to say something about the overall effects of a program as well as particular effects in subunits of the organization. The two are not mutually exclusive.

### Triangulation

In designing naturalistic evaluation, the strategy of triangulation is important. Triangulation is defined (Denzin, 1970, p. 297) as the "combination of methodologies in the study of the same phenomena." Denzin (1970) identifies four types of triangulation: (1) investigator (multiple evaluators investigating the same program); (2) data source (use of as many data sources as possible to understand events being analyzed); (3) data collection methods

### Exhibit 6-1. Unit of Analysis--An Educational Conference Evaluation

I was asked by members of the conference committee to design and conduct an evaluation of the lived experience of participants at the 1993 American Educational Research Association Annual Meeting (AERA). This is the annual meeting of a large organization of educational researchers and evaluators with diverse content and research interests.

Prior to making any design decisions, I asked the committee chairperson why they wanted the study done, how they would use the data, and whether they wanted data from persons in all levels and divisions of the organization. I learned that what it was they wanted me to be able to say something about when I finished was what different people's experiences of the conference were on a daily basis so that they could consider making changes which would enhance the experience for all participants in future years.

Based on that understanding, I knew that I needed an ethnographic evaluation design to be able to hear different persons' feelings and reflections as well as observe their behaviors and interactions of the lived experience within the larger culture of the organization on a daily basis. I made a decision to select 5 interviewers and target about 15 participants to obtain in-depth information and thick description (the *meanings* that different events at the conference had for them) rather than to use a larger sample and different data collection techniques for increased breadth of understanding. I also carefully sampled persons to get diversity in research approach (qualitative or quantitative), geographic location, gender, academic status, and years of membership in AERA so that I would be able to say something about the lived experiences of different levels of persons in the organization.

(within-method and across-method); and (4) theoretical (approaching data with multiple theoretical perspectives and hypotheses). Multiple triangulation is the use of all forms of triangulation. The two most frequently used types of triangulation in evaluation are data source and data collection methods. Data source triangulation includes data collection from different levels of persons, different times, and different places at the site. Data collection methods triangulation in evaluation includes the use of different forms of the same approach as well as different techniques of data collection such as interviews, observations, open-ended questionnaires, surveys, and so on. Denzin (1970) asserts that use of multiple methods of data collection reduces

threats to validity in that weaknesses in one method are offset by strength of another. Multisite investigations are another way evaluators can triangulate the design.

Evaluators believe that triangulation will result in corroborative data across sources, methods, or sites. Triangulation is commonly perceived as a strategy for enhancing validity of research findings. Researchers (Caracelli & Greene, 1993) assert that triangulation seeks convergence, corroboration, and correspondence of results across different methods. Miles and Huberman (1984, p. 234) assert that "triangulation is supposed to support a finding by showing that independent measures of it agree with or, at least, don't contradict it." Mathison (1988, p. 13) notes that historically it is seen as "a strategy that will aid in the elimination of bias and allow the dismissal of plausible rival explanations such that a truthful proposition about some social phenomenon can be made." Mathison (1988, p. 17) argues, "More realistically, we end up with data that occasionally converge, but frequently are inconsistent and even contradictory." This understanding of the result of triangulation places the burden on the evaluator of collecting data which explain *why* data are different or contradictory from different data sources about the same social phenomenon. When evaluators use across-method data collection triangulation as described in Exhibit 6-2, it is important to note comments made by participants about the different data collection methods which might provide insights for valid interpretation of data and the ability of the evaluators to explain differences or conflicts in data from the same data sources and contexts.

### Mixed-Method Designs

Mixed-method designs have been defined as those that include at least one quantitative method and one qualitative method where neither type of method is inherently linked to a particular inquiry paradigm or philosophy (Caracelli & Greene, 1993). A mixed-method design using investigator triangulation, where the evaluation team consists of both qualitative and quantitative evaluators committed to their inquiry paradigm and philosophy, is a particularly strong design, however. Evaluators bring extensive training, expertise, and experience in their particular paradigm and data collection approach to inform different aspects of the evaluation. This approach addresses concerns raised by Guba and Lincoln (1988) that internal consistency of each paradigm would be violated by mixing different inquiry approaches. It can be argued (Brown, 1992) that different and important understandings can emerge by triangulating qualitative and quantitative evaluation methods using investigators who are strong in each approach.

**Exhibit 6-2. Triangulation: Generating Understanding from Data Obtained from Different Methods**

As a qualitative evaluator, I worked on an evaluation team with a quantitative evaluator to investigate the impact of a climate improvement program in an elementary school. At the end of the second year of the program, we asked all teachers in the school to meet in the media center to provide information on two instruments. The quantitative evaluator administered a standardized Likert-scale response survey. I administered an open-ended questionnaire with non-leading questions I had written to allow teachers to write about the meaning and impact of the climate improvement program on them, their students, and the school. Questions on the open-ended questionnaire included things such as the following: 1) Describe ways the climate improvement program has impacted you and your students; 2) What were the most positive aspects of the project for you personally and professionally? 3) What were the most negative aspects of the project for you personally and professionally? We got different results from the same group of people about the same phenomenon using across-method triangulation. Comments made by some of the teachers as they left the media center provided the "why" for us. Each of them that reflected on the process indicated that the responses they wrote for the open-ended questionnaire "really told it like it is" and the responses on the survey did not because it "didn't ask the right questions" and "didn't provide enough responses to select from."

In mixed-method designs such as the one described in Exhibit 6-3, data from each paradigm are analyzed independently. Quantitative data are numerical and are analyzed statistically. Qualitative data are generally narrative (but sometimes numerical for descriptive purposes) and are analyzed using a strategy like constant comparative analysis or phenomenological analysis which allow for emerging categories and relations among categories to be generated from participant data. There is no need to use strategies to artificially numerically code and transform rich narrative data into numerical form for analysis that would be comparable to quantitative data analysis. That undermines the basic reasons for conducting rigorous in-depth qualitative evaluation. Analyzed and interpreted data from each approach are examined and compared by all evaluators to generate a broader understanding of the impact of the program being evaluated.



**Exhibit 6-3. Mixed-Method Evaluation of an Innovative Preschool Program for At-risk Children**

I was the qualitative evaluator for an innovative preschool program for at-risk children where triangulation of investigators was a planned part of the evaluation design. Statistical comparisons of pre- and post-implementation data from a variety of standardized instruments clearly demonstrated the positive impact of the program on understandings and skills development of the preschool children. From a different paradigm, the qualitative paradigm, findings from observations, interviews, and videotape analysis revealed the positive impact of the program on the teaching--learning situation, parent involvement, student--parent interactions, community involvement, and on-location learning at sites other than the classroom. Evaluation reflecting philosophical underpinnings and data collection methods of either paradigm alone could not have resulted in the rich diversity of understanding that resulted from the mixed-method investigator triangulation design.

**Ethical Considerations**

All evaluators must consider the ethical implications of their work. Qualitative evaluators must exercise extreme caution in detailing every aspect of the social system and social hierarchy of an organization prior to gaining entry in order to avoid as many ethical blunders as possible. Qualitative evaluators, and particularly ethnographic evaluators, will spend a great deal of time in the site interacting with and observing persons. Every conversation for the ethnographic evaluator is data collection. Because the evaluator is an outsider, persons at different levels begin to reveal aspects of the life and culture at the site which may bias the evaluator in data collection and interpretation.

More important, however, the information itself may present ethical dilemmas where the evaluator has to decide whether to reveal information to stakeholders relative to anticipated injury, social loss to participants who provide the information, or other potential changes in the social structure or organization that would not occur if s/he kept the information confidential. If a participant asks the evaluator not to use information and the evaluator agrees, then whatever the participant shares with the evaluator, even if it is critical for an accurate evaluation, must be excluded from fieldnotes and the final report. If the evaluator indicates that participant information will remain confidential, then fieldnotes, interview transcripts, and the final evaluation report must reflect that promise. Sometimes (see Exhibit 6-4) evaluators

inadvertently break the promise of anonymity by the way data are reported. This can be avoided by carefully masking sources and using member checks before submitting reports.

#### **Exhibit 6-4. Ethical Decisions--Confidentiality Agreement Broken**

In conducting individual interviews for the second year of an ongoing project, I began the interview with the first site manager in my usual manner stating, "I will be interviewing the manager of each site, but the information you share about the impact of the project at your site will remain totally confidential in that I will not use your name in my notes and neither you nor your site will be recognizable in the final evaluation report." The first interviewee sat quietly, answered questions politely, and left. I was puzzled in that the rapport I usually quickly establish in interview situations never materialized.

The second manager, after hearing my promise, leaned back in his chair laughing loudly and said, "Yea"? Well that's just what the last interviewer said last year and when the boss got the report, we were labeled Site #1, Site #2, Site #3, and so on, and he and everyone else knew by the clear descriptions exactly who had said what and what was going on at each site. I'm not going to tell you anything that isn't common knowledge around here and none of the rest are either."

I had a very difficult time convincing those managers that I would not write the evaluation report in the same way. I was finally able to gain the trust of all of the participants, but at great cost in time in each interview as I had to explain what I had learned about the last year's evaluation and how I would function differently. Then I assured each manager that I would write the evaluation report, send a copy to each of them to edit as they felt it needed to be to protect individual identities, and only after that, send the final evaluation report to their boss. They agreed to that process. I received no suggestions for revisions of the final evaluation, but only approvals to send it forward.

Being in the natural setting and "living the program" with the participants, the qualitative evaluator will be privy to conversations about controversial situations in the site, many of which do not have anything to do with the project being evaluated. The best advice in such situations is to listen and respond normally as one would in any conversation, but do not act on anything shared in informal conversational situations unless it relates to the project. Personal information about participants or other non-project events should not be recorded or repeated by the evaluator. The evaluator's job does

not include righting all wrongs or instituting changes that appear necessary. The goal of qualitative evaluation is simply to determine the impact of the program on participants at the site by understanding the meanings it has to them; why it has those meanings; and how it affects behavior, actions, and interactions in that context.

Participants should be interviewed in the least threatening circumstances and locations at the site. Teachers, for example, at one school forewarned the evaluator that they would casually walk away if a particular female colleague came into the area because they did not want her to think they were "telling [the evaluator] what is really happening in the project" and make the project director and principal angry with them. Recorded data (fieldnotes or tapes) should be destroyed as soon as data have been used to generate the evaluation report. Because of the personal nature of data collection in participant observation evaluation, participants feel particularly betrayed when they reap negative and unexpected impacts of an evaluation effort after providing the evaluator with ongoing information for an extended time.

Participants at a site should be given an opportunity to hear about the evaluation design and their roles in it initially and then be provided with an opportunity to give written consent to participate. Qualitative evaluators should not share any participant's perspectives with others at the site. Fieldnotes and interview transcripts should remain confidential. Analyzed and interpreted data, on the other hand, can and should be shared with participants as member checks (Guba & Lincoln, 1989) to determine whether the final understandings make sense to them and make sure that the evaluator "got it right."

### **DATA COLLECTION METHODS**

A variety of data collection methods are associated with qualitative evaluation designs. These include participant observation, observation, interviews, open-ended questionnaires, and document collection. Characteristics shared by those methods are that they are unobtrusive, inductive, labor and time intensive, and generally result in narrative data. The goal of using these types of data collection methods is to generate data from the perspectives of the participants in programs being evaluated.

#### **Participant Observation**

Participant observation is the most common data collection method in qualitative evaluation. The role assumed by the evaluator falls on a continuum (Gold, 1958) from total researcher to total participant. Most qualitative evaluators assume a role more toward the total researcher end of the continuum. This means that the evaluator does not assume an active role within the social group in the ongoing program. Taking the role of total

evaluator requires a period of negotiating a trust relationship with participants at the program site. Gaining entry, establishing a trust relationship, negotiating reciprocity, finding an unobtrusive niche, and meeting persons at different levels in the program are important steps for successful qualitative evaluation.

A participant observer participates in and observes as much of the social interaction relative to the program in the natural setting as possible. Observation is unstructured, holistic, and constant in the setting. Participant actions, interactions, and responses to programs guide observations. A "verbal photograph" (what is happening here) of the social actions and interaction can be recorded in fieldnotes, audio tapes, or videotapes. The advantage of fieldnotes recorded by hand or computer is that the lived experience does not have to be "lived again" as it does when the evaluator watches or listens to tapes of events. Each entry in a field notebook should be contextualized, have the date and time, and include a brief description of the event and persons involved in it. The disadvantage of recording events in fieldnotes is that interactions are missed when the evaluator is writing. An advantage of videotapes of events is that the event can be observed multiple times to obtain different types of information (verbal interactions, nonverbal interactions, and context clues). In addition, participants can watch segments of videotapes with the evaluator and provide interpretations of interactions from the insider's perspective.

Fieldnotes often include observer comments and reactions to things observed. Fieldnotes are for the exclusive use of the evaluator and are not shared with participants. Daily ongoing systematic analysis of fieldnotes provides a guide for further observation and interviewing needs throughout the evaluation. The ongoing analysis allows the evaluator to identify phenomena and events that are clearly understood, missing, and incomplete. Changes in observation strategies are based on these understandings.

### Interviewing

The purpose of interviewing in qualitative evaluation is to find out the meaning of the program to participants. Interview formats can vary on a continuum from highly structured evaluator directed question and response guides to informal conversations whose focus and direction are directed by participants. Participant observation always includes conversational interviews because of the level of interaction between the evaluator and participants. The selection of interview format is determined by the type of information that is desired, the amount of time available to the evaluator to collect data, and the level of comparability of findings that is desired. The less structured interview formats require more time and are less comparable, but they allow participants to discuss issues and concerns that are of utmost importance to

them. Evaluators must carefully consider the tradeoffs when selecting interview formats.

Interviews can be with individuals or groups of participants. Each approach has advantages, and both can be included in a single evaluation design. Focus group interviewing is a form of qualitative data collection in which the evaluator functions as discussion facilitator for a small group of participants and relies on interaction within the group to provide insights about topics proposed by the evaluator. Krueger (1988) argues that focus group interviews can provide vital information on the impact of programs on participants. Morgan (1988) explores the advantages and disadvantages of focus group interviews.

Advantages of focus groups are:

- They are relatively easy to conduct.
- They require less time than multiple individual interviews.
- They provide the opportunity to collect data from group interaction.
- They provide an opportunity for group discussion opinion formation of researcher-generated topics.

Weaknesses of focus group interviews are:

- They are not conducted in the naturalistic setting.
- It is impossible to discern individuals' perspectives.
- The degree to which the presence of the evaluator and other participants affects responses of any individual cannot be determined.
- Comparison of data across focus groups is difficult because group interaction determines the direction or focus of discussion.
- Fewer questions can be asked because more interviewees are involved.

In addition to the format, the wording and sequencing of questions affect interviewee responses. Interview questions in qualitative evaluation should be singular, clearly worded, nonleading, and open-ended (Patton, 1990, pp. 277-368).

One of the most important goals of interviewing in evaluation is to find out *why* different individuals or different levels of persons construct different meaning about a program. In other words, in order to be able to say something about the reasons for different or conflicting findings about a program, data need to be generated that account for those differences in perspective or meaning. One of the most *ineffective* question formats is to ask "Why?" after other questions. Role playing and simulation questions or

questions asking for descriptions or examples are superior techniques for finding out why participants function the way they do or have the perspectives they share about a program.

Interviewing as many participants as possible in different contexts and across the times throughout the program will provide an understanding of evolving perspectives. Key informants are special people in the social context with whom the evaluator spends more time than with other participants. The key informant provides insights and insider interpretations that the evaluator may not be able to access as an outsider to the group and situation. Key informants are selected because they may be particularly well informed about the program, may be available to the evaluator, may have played a key role in helping the evaluator gain access to the site, or other characteristics that make her/him special and different from other participants. Selection of a key informant who is peripheral in the social structure or who is viewed negatively by some or all of the program participants can be detrimental to the evaluation process.

### Questionnaires

Carefully constructed, open-ended questionnaires serve the same purpose as interviews in that they help the evaluator can "get inside the head" of participants to find out their perspectives of the program (see Exhibit 6-5.)

Questionnaires require less time to administer than interviews so comparable data can be collected from many more participants. (See Chapter 7 for additional discussions of questionnaire and opinionnaire design and use.) Cautions in composing questions for an open-ended questionnaire include avoiding leading questions, writing clear questions, providing enough space for responses, and carefully arranging questions so that a response to one will not affect the response to subsequent questions. Questionnaires should be as concise as possible to ensure the greatest number and highest quality of responses. Pose only those questions whose responses will allow you to be able to say something about the phenomena you want to say something about as defined by your selected unit of analysis.

Some issues need to be addressed when using open-ended questionnaires.

1. The first issue is whether questionnaires should be administered in the program context or mailed to participants. Mailed questionnaires usually result in a relatively low response. On the other hand, there may be time or human presence constraints in the context that could affect the way participants respond to a questionnaire.

2. The second issue is whether questionnaires should be confidential or anonymous. If comparison of data by individual participant with data from other methods is essential, then participants would be required to supply their name or some different form of identification which would allow them to remain anonymous. When participants are required to put their names on questionnaires, the content of responses may be affected if participants feel that they or their job security is threatened by persons at different levels within the organization if they respond honestly.
3. The third issue is whether questionnaire data will be analyzed by person across questions or by question across persons. If data are compared by question across persons, the identity of individuals is generally easily masked. When data are analyzed by individual or subgroups of individuals, identities of respondents can often be identified by persons within the context.

Decisions about each of these issues must be made after consideration of each evaluation situation and context to determine which approaches will result in the most valid information and the highest response rate.

### **DATA ANALYSIS**

Analysis of qualitative data is an ongoing cyclical process that consists of synthesizing information across data sources and data collection methods. The analysis process is generative in that hypotheses are not tested but generated from participant data. There are different approaches to qualitative data analysis, and each addresses different evaluator needs relative to the types of data collected and evaluation questions asked.

Most qualitative data are in narrative form, but often qualitative evaluators have numerical data that are presented as descriptive statistics. Examples of numerical data in qualitative evaluations might be proportions of different categories of persons participating in a program, proportions of time participants are engaged in specific activities on a daily basis, group sizes, and other such frequency counts. Stakeholders want to know how representative claims in evaluation reports are, but this is not an argument that supports the transformation of rich narrative qualitative data into simple frequency counts.

#### Qualitative Data Analysis Strategies

Qualitative data analysis is inductive. Evaluators engaged in interpretive inquiry generally begin with rich data from a variety of data sources and methods to determine "what is in the data" rather than beginning with a theory or hypothesis to test. Identification of categories or themes in the data is followed by establishing relations among categories and then seeking further

evidence to support categories and relationships in the field setting. In some cases, qualitative evaluators will, however, begin with a theory to test and use analytic induction to analyze data.

**Exhibit 6-5. Example of an Open-ended Questionnaire**

TEACHER QUESTIONNAIRE ON  
STAFF DEVELOPMENT EXPERIENCE

Instructions: Please take the time to read carefully and respond honestly to each question. We sincerely want to know about your experience in the Project 2061 staff development workshop and value your suggestions for improvements. Thank you.

1. Describe any ways that you think differently about the way children learn or the way children learn science as a result of experiences or articles given to you in this staff development.
2. Describe ways that you will teach science differently next year because of things you read, learned, or experienced during this Project 2061 staff development.
3. How would you describe what science is to a group of students in your school?
4. What was the most positive aspect of the Project 2061 science workshop for you? What about it was positive for you?
5. What was the most negative aspect of the Project 2061 science workshop for you? What about it was negative for you?
6. If you were going to conduct a similar Project 2061 science workshop for teachers, what changes would make to improve teachers' experiences?

Phenomenological Analysis. The goal of phenomenological analysis (Hycner, 1985) of narrative data, particularly interview data, is to understand the phenomenon or program in its own right and not from the perspective of the researcher. The evaluator brackets or suspends her/his own meanings and interpretations as much as possible and allows meaning to emerge from the data (interviews, questionnaires, open-ended surveys, and documents) that have been generated by participants. The evaluator as analyst delineates units of meaning in participant data relevant to the evaluation questions and established relations among units generated in different data sources and data



collection methods. The clusters of units of meaning are themes that address the evaluation questions.

Content Analysis. Content analysis is a well known method for analyzing documents and written communication. Documents are often produced at a program site without guidance from the evaluator and for different reasons other than evaluation. Documents, however, can be a good source of information about program implementation and interpretation by participants. Content analysis is defined by Holsti (1969, p. 14) as "any technique for making inferences by objectively and systematically identifying specified characteristics of messages." Guba and Lincoln (1981) make a case for qualitative content analysis, explaining that frequency counts are not necessarily associated with the importance of assertions in documents. Qualitative content analysis includes generation of categories from the data which are relevant to the purposes of the evaluation. Evaluator-generated rules for categorization, demonstration of representativeness of categories, relations among categories, and definitions of categories from participant perspectives are important outcomes of content analysis.

Analytic Induction. Qualitative evaluators beginning with a theory to test about a program in a particular setting would probably analyze data using analytic induction. Rather than starting with holistic observation and interviewing, cases would be selected using specific criteria in the setting to test the theory. As data saturation (finding no new or different cases) occurs, the evaluator stops collecting data and presents the evidence to support the theory.

Constant Comparative Analysis. Constant comparative analysis (Glaser & Strauss, 1967; Strauss, 1987) is an approach to analysis that results in grounded theory. Analysis is ongoing throughout data collection. As data are displayed and reduced into categories of meaning and relations among categories, hypotheses are proposed to account for social meaning and interaction that emerge in the data. Through theoretical sampling, the evaluator is guided in data to collect, data sources to approach, and data collection methods to use. A process of writing theoretical and methodological memos, or notes about ongoing insights, informs the interrelated data collection and analysis process.

### Presentation of Evaluation Data

The key to presenting qualitative or ethnographic evaluation findings effectively is the idea of contextualization or interpretation of behavior, interactions, and constructed meanings within the context or culture in which they were collected. Historically, qualitative evaluators have not been concerned about generalizability of findings, but rather are concerned about presenting an accurate and holistic description of the immediate and larger

contexts in which findings are generated. The strength of qualitative evaluation is the generation of in-depth understanding of the process, social interaction, participants' perspectives, and meanings constructed by participants in programs being evaluated.

Multisite qualitative evaluation designs are a way to enhance generalizability of findings (Herriott & Firestone, 1983). Multisite evaluations of the same program in dissimilar contexts provides greater generalizability than single site designs or multisite designs where contexts are similar (Kennedy, 1979; Sinacore & Turpin, 1991). Firestone (1993) discusses issues related to generalizability of qualitative research and presents an approach using Boolean algebra to compare large numbers of cases systematically. Generalizability, however, is not the goal of qualitative research and evaluation. Generalizability of qualitative findings depends on the degree to which other situations, programs, and participants are similar to those described in the evaluation report. Thus, the primary responsibility for generalizability rests with the evaluator as the program, context, social structure, and participants are clearly, systematically, and holistically described relative to program related findings.

Qualitative evaluation findings may be shared throughout an evaluation to guide changes in programs, presented in a final report at the end of the piloting of a program or in both ways. In a multiphase or multiyear program in which an evaluation report must be present at the end of each segment, or when the goal is formative evaluation, it is important to remain cognizant of the possible negative impact of interim reports and responses to those reports by stakeholders on participants at all levels of the organization.

Many stakeholders in programs have constructed misunderstandings about the nature of naturalistic and interpretive inquiry without having any formal training in it. Although they know neither the types of questions which are best answered using naturalistic evaluation nor the methodology of interpretive inquiry, they form very strong beliefs about the nature of findings and value of this approach to knowing. It is critical to keep this in mind for the duration of a qualitative evaluation, knowing that there are many ways in which the evaluation process or stakeholders' responses to it might negatively affect one or more participants (see Exhibit 6-6.)

The specific and important role that interpretive inquiry plays in evaluation was considered in this chapter. Interpretive inquiry in evaluation can be in the form of ethnographic evaluation, qualitative evaluation, or naturalistic evaluation. These approaches to evaluation are conducted in the natural setting with an emphasis on understanding the program impact from the perspectives of all levels of persons in the organization. When the goal of evaluation is to find out the impact of a program on social interaction, the

**Exhibit 6-6. Demoralized Participants--A High School Vocational Education Project**

At the end of the first year of a four-year high school vocational education project to enhance basic skills of vocational students by having vocational and academic teachers plan together, I was required to present a qualitative evaluation report to the project director, school principal, county superintendent, participating teachers, and project validation onsite team members. My report was based on a vast amount of very rich data from teachers and students and indicated a number of very positive outcomes of the project for teachers and students in the school. The quantitative evaluator indicated that there were not sufficient test data to make an evaluation that early. After both reports were presented, the superintendent asserted that the qualitative data could be "manipulated to show anything," and the chairperson of the onsite team proclaimed that while the qualitative data were interesting, that the "jury would remain out on the effect of the project until the hard data [test scores] were in."

Teachers had cooperated totally with self-reflection procedures required in the qualitative design. They kept weekly logs of events and experiences, they agreed to multiple interviews with me, they allowed me to interview students and administer open-ended questionnaires to students, and they provided me with all documentation relative to the project. It had been time consuming, but they had felt it was valuable as they began to use their own data to make decisions about needed changes to make the project more effective.

At the end of the reporting meeting I left with a large group of project teachers. All teacher comments went something like, "Forget it! We won't do any more logs or anything. We thought we were doing good things, but they don't value what we've done at all. It simply goes back to test scores again--are they significantly better or not! They don't even care if kids are getting a better education and we are doing a better job. They don't care what we think or what we do." Needless to say, qualitative data collection the last three years of the project was very difficult. Teachers understood the need for the process for themselves and the good of the project, but they were very discouraged by the responses of the administrators and outside validation team members to the positive information that had been presented.

construction of social meaning, individuals in a social setting, or to find out what is actually being done at a site that is piloting or implementing a program, an interpretive or qualitative paradigm provides the epistemological and methodological basis for providing understanding.

Evaluators trained in the use of qualitative data collection and analysis methods are the best persons to design and conduct qualitative evaluation. Schooling in the philosophical and theoretical underpinnings of qualitative inquiry allows the evaluator to make informed decisions throughout the evaluation concerning ethical issues in the field, the role of subjectivity, evaluator roles, and use and interpretation of multiple perspectives and triangulation. Qualitative and ethnographic evaluation is fieldwork based and includes data collection methods of observation, interviewing, and document collection. Generally, questions and hypotheses are generated from data collected about the program or phenomenon being evaluated. In some cases, however, evaluation questions and data categories are established prior to fieldwork and data are collected specifically to test those questions.

Qualitative evaluation data are analyzed in an ongoing and interactive manner with data collection so that findings can guide hypothesis generation and testing. The data collection method and evaluation questions serve as a guide to selection of an appropriate data analysis strategy. Systematic and rigorous qualitative data collection with ongoing data analysis and interpretation can provide formative and summative participant-based information for enhanced stakeholders' decision making.

### COGITATIONS

1. Which types of evaluation goals would best be addressed by using qualitative evaluation? Ethnographic evaluation?
2. What are the characteristics of the qualitative paradigm that inform qualitative and ethnographic evaluation?
3. How does the concept of multiple perspectives of reality support Mathison's assertion that triangulation often results in convergence, inconsistency, and contradiction?
4. What are the strengths and weaknesses of different data collection methods in qualitative evaluation, and what are the best ways to offset those weaknesses?
5. What issues must be addressed when designing qualitative evaluation?
6. What are different approaches to analyzing qualitative evaluation data? What are the advantages and disadvantages of each approach?
7. What meaning does generalizability have in qualitative evaluation?

### SUGGESTED READINGS

- Guba, E.G., & Lincoln, Y.S. (1981). *Effective evaluation*. San Francisco: Jossey-Bass.
- Guba, E.G. & Lincoln, Y.S. (1989). *Fourth generation evaluation*. Newbury Park: Sage.
- Krueger, R.A. (1988). *Focus groups: A practical guide for applied research*. Newbury Park, CA: Sage.
- Patton, M.Q. (1990). *Qualitative evaluation and research methods*. (2nd ed.) Newbury Park, CA: Sage.
- Strauss, A.L. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University Press.