CHAPTER 5

APPLYING QUALITATIVE EVALUATION METHODS

INTRODUCTION ●

This chapter introduces a different approach to program evaluation—one that has emerged in parallel to the more structured, quantitative approach that has been elaborated in Chapters 2, 3, and 4. This chapter will show how **qualitative evaluation methods** can be incorporated into the range of options' available to evaluators and their clients, and will offer some comparisons between the two approaches. In general, qualitative approaches, such as interviews and **focus group** discussions, are more open-ended than quantitative methods, and are most valuable in collecting and analyzing data that do not readily reduce into numbers. Qualitative methods are particularly useful for exploratory research and participatory, or empowerment, research. Empowerment research (see Fetterman, 1996) involves significant collaboration between the evaluator and stakeholders during most or all of the steps in the evaluation process, from the planning and design to the final interpretation and recommendations. Chapter 11 includes further information on empowerment evaluation in its discussion of the relationship between program evaluation and performance management.

It is worth reiterating that the key issues in deciding on which method or methods to use for any evaluation are the context of the situation and the evaluation questions that need to be addressed. Qualitative methods can be used in various stages of an evaluation:

- Determining the focus of the evaluation
- Evaluating the implementation or the process of a program
- Determining improvements and changes to a program

To introduce qualitative evaluation methods, it is important to first elaborate on the diversity of approaches even within the theory and practice of qualitative evaluation. Qualitative evaluation approaches differ from each other on at least two important fronts: their philosophical beliefs about how and what we can know about the kinds of situations evaluators typically face (these are called **epistemological beliefs**); and their methodologies, that is, the ways that evaluations are organized and conducted. In this chapter, we will learn about some of the key philosophical differences among qualitative evaluation approaches, but will spend more time focusing on the ways that qualitative methods can be used.

● COMPARING AND CONTRASTING DIFFERENT APPROACHES TO QUALITATIVE EVALUATION

When qualitative evaluation approaches emerged as alternatives to the then-dominant **social scientific approach** to evaluation in the 1970s, proponents of these new ways of evaluating programs were a part of a much broader movement to re-make the foundations and the practice of social research. Qualitative research has a long history, particularly in disciplines like anthropology and sociology, and there have been important changes over time in the ways that qualitative researchers see their enterprise. There is more diversity within qualitative evaluation approaches than within quantitative approaches:

> A significant difference between qualitative and quantitative methods is that, while the latter have established a working philosophical consensus, the former have not. This means that quantitative researchers can treat methodology as a technical matter. The best solution is one which most effectively and efficiently solves a given problem. The same is not true for qualitative research where proposed solutions to methodological problems are inextricably linked to philosophical assumptions and what counts as an appropriate solution from one position is fatally flawed from another. (Murphy, Dingwall, Greatbatch, et al., 1998, p. 58)

Denzin and Lincoln (2000) summarize the history of qualitative research in their introduction to the *Handbook of Qualitative Research.* They offer an interpretation of the history of qualitative research in North America as comprising seven moments, beginning with traditional anthropological research (1800s to about 1950), characterized by lone anthropologists spending time in other cultures and then rendering their findings in "objective" accounts of the values, beliefs and behaviors of the natives. In their definition of qualitative research Denzin and Lincoln (2000) include the following:

> Qualitative research is a situated activity that locates the observer in the world . . . qualitative researchers study things in their natural settings, attempting to make sense of, or to interpret, phenomena in terms of the meanings people bring to them. (p. 3)

and later continue:

> Qualitative researchers stress the socially constructed nature of reality, the intimate relationship between the researcher and what is studied,

> and the situational constraints that shape inquiry. Such researchers emphasize the value-laden nature of enquiry. They seek answers to questions that stress *how* social experience is created and given meaning. In contrast, quantitative studies emphasize the measurement and analysis of causal relationships between variables, not processes. (p. 8)

Understanding the Issue of Paradigms

In the field of program evaluation, academics and practitioners in the 1970s were increasingly under pressure to justify the then-dominant social science-based model as a way of thinking about and conducting evaluations. Questions about the relevance and usefulness of highly structured evaluations (often experiments) were being raised by clients and by academics alike. An alternative paradigm was emerging, based on different assumptions, different ways of gathering information, different ways of interpreting that information, and finally, different ways of reporting evaluation findings and conclusions.

On the whole, the qualitative research approach embraces a different view of research than the **positivist** "rational" approach traditionally taken by most quantitative researchers. Thomas S. Kuhn (1962), in his revolutionary book *The Structure of Scientific Revolutions,* asserted that when scientists "discover" a new way of looking at phenomena, they literally *see the world in a different way.* He developed and popularized the notion of a **paradigm**, a self-contained perceptual and theoretical structure akin to a belief system. Although Kuhn was describing the change in world view that happened in physics when Einstein's relativity theory began its ascendancy at the turn of the 20th century, he used language and examples that invited generalizing to other fields. In fact, because his book was written in a nontechnical way, it became a major contribution to the widespread and continuing process of questioning the foundations of our knowledge in the social sciences and humanities.

Paradigms, for Kuhn, were at least partly **incommensurable.** That is, adherence to one paradigm, and its attendant way of seeing the world, would not be translatable into a different paradigm. Proponents of dissimilar paradigms would experience an inability to communicate with their counterparts. They would talk past each other, because they would use different words and literally see different things even when they were pointing to the same object.

Norwood Russell Hanson (1958) illustrated this problem with a series of optical illusions and puzzles. A well-known one is reproduced here to make several points about seeing. Figure 5.1 is a line drawing of a person. Look

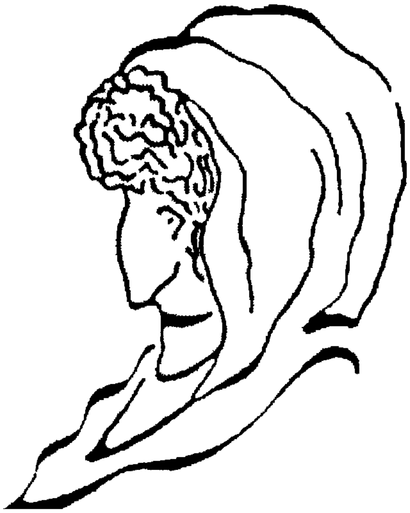

Figure 5.1 The Old Woman or the Young Woman?

carefully at the drawing and see if you can discern the old woman's face. She is facing the left side of the page, her nose is prominent, her lips are a line, and she is wearing a hood that exposes the hair on her forehead.

Now, look at the same drawing again and see if you can discern the young woman's face. She is facing away from you, and is showing you the left side of her face. You cannot see her nose, but you can see her left ear. Her jaw and her hair are visible as is her necklace. A hood that loosely covers her head exposes the left side of her face.

Generally, people find it easier to "see" the old woman. But once you "see" the young woman, she will be there every time you look for her.

The point of Figure 5.1 is to show you that the same information can be interpreted in different ways. Further, the patterns of information are incommensurable with each other. When you are seeing the old woman, you are not (at that moment) seeing the young woman. Kuhn was arguing that scientists, too, can "see" the same information and interpret it differently.

In the qualitative methods arena, one key paradigm, often called the **constructivist** paradigm, has been developed and articulated by a number of central contributors. Egon Guba and Yvonna Lincoln (2001) are two well-known scholars who have made major contributions to constructivist research approaches. In explaining relativism, one of the fundamental assumptions of the constructivist paradigm, they state that "human (semiotic) sense-making that organizes experience so as to render it into apparently comprehensible, understandable, and explainable form, is *an act of construal and is independent of any foundational reality.* Under relativism there can be no 'objective' truth" (p. 1, original underlining).

Guba and Lincoln are among those who have stated a belief that the "scientific" approach to program evaluation and the constructivist approach are fundamentally different, and at least partially incommensurable, paradigms:

> It is not appropriate to "mix and match" paradigms in conducting an evaluation, for example, utilizing both scientific (positivist) and constructivist propositions within the same study. This is not a call for "purity" nor is it intended to be exclusionary. It is simply a caveat that mixing paradigms may well result in nonsense approaches and conclusions. (Guba & Lincoln, 2001, p. 2)

However, other influential evaluators, including Michael Patton, while acknowledging the continued existence of paradigmatic differences between social scientific evaluators and some qualitative research approaches, have argued that the philosophical differences have been at least substantially submerged by more practical concerns:

> The trends and factors just reviewed suggest that the paradigms debate has withered substantially. The focus has shifted to methodological appropriateness rather than orthodoxy, methodological creativity rather than rigid adherence to a paradigm, and methodological flexibility rather than conformity to a narrow set of rules. (Patton, 1997, p. 295)

For Patton, pragmatism makes it possible to overcome the differences between paradigms:

> I believe that the flexible, responsive evaluator can shift back and forth between paradigms within a single evaluation setting. In doing so, such a flexible and open evaluator can view the same data from the perspective of each paradigm and can help adherents of either paradigm interpret data in more than one way. (Patton, 1997, p. 296)

Patton's view is by no means universal, however. The recent discussions about the primacy of randomized control trials (RCTs) within the evaluation community (Scriven, in progress, unpublished) are connected with the ongoing debate about the relative merits of qualitative and quantitative evaluation approaches.

The Pragmatic Approach

In evaluation, although there continue to be debates that stem from differences in epistemological beliefs and methodological approaches, there is a general movement toward Patton's more pragmatic pluralism. Even in a textbook that has been considered to be a benchmark for **positivist** and **post-positivist** approaches to evaluation (Cook & Campbell, 1979), the more recent edition (Shadish, Cook, & Campbell, 2002) mentions the value of qualitative approaches as complements to the experimental and quasi-experimental approaches that comprise the book.

Mark, Henry, and Julnes (2000) have offered a theoretical approach to evaluation that they claim will settle the qualitative/quantitative dispute in the profession. Their approach relies on combining what they call "**sense-making**" with commonsense realism:

> Neither qualitative nor quantitative methods are superior in all evaluation situations. Each inquiry mode corresponds to a functional aspect of sensemaking and each can be addressed through qualitative and quantitative methods. More generally, commonsense realism, integrated with sensemaking, offers a potent grounding for a lasting peace following the paradigm wars. (p. 335)

The paradigm debate is not dead yet. But its role in the practice of program evaluation has diminished considerably. Most program evaluators have taken the position that qualitative and quantitative methods do not carry the freight of different philosophical traditions, or if they do, methodological pluralism is the solution of the day. Ernest House (1994), for example, has argued that the quantitative-qualitative dispute is dated and has stated that he does not believe that the two methods represent distinct paradigms that incorporate incommensurate worldviews. The two methods can be applied regardless of whether one believes that we share the same reality, or that each person has a reality that is ultimately known only to the perceiver.

● QUALITATIVE EVALUATION METHODS: SOME BASICS

What is qualitative evaluation? How is it distinguished from other forms of program evaluation? How do qualitative evaluators do their work?

These questions are practical ones, and the main focus of this section will be to offer some answers to them. It is worth saying, however, that qualitative evaluation methods have developed in many different ways and that there are a number of different textbooks that offer evaluators ways to design, conduct, and interpret evaluations that rely on qualitative data (for example Denzin & Lincoln, 2000; Patton, 1997). Patton, in the Evaluation Checklists Project (The Evaluation Center, 2001) maintains "Qualitative methods are often used in evaluations because they tell the program's story by capturing and communicating the participant's stories." They normally encompass interviews, focus groups, narrative data, field notes from observations, and other written documentation. Often, the sample size is quite small. In contrast, quantitative evaluations use numbers gathered from measures over comparatively large samples, and use statistical procedures for describing and generalizing the patterns between and among variables.

An evaluation may be entirely conducted using a qualitative approach, but it will depend on the context and needs of the evaluation. Sometimes initial qualitative exploratory work is followed by a quantitative approach, particularly when an evaluator is developing survey questions. Developing logic models is a qualitative process that relies on interpreting documents, interviewing stakeholders, and putting together a visual representation of a program. Sometimes qualitative findings are collected and/or presented along with quantitative data, such as that gathered from a survey with both closed-ended and open-ended questions. Finally, qualitative research sometimes occurs *after* quantitative research has been completed, such as when an organization is determining how to follow up on survey results that indicate a program needs changes.

There are strong practical reasons to view qualitative evaluation methods as complementary to **quantitative methods**. Indeed, as Reichardt and Rallis (1994) and many others have argued, using two methods can be better than one. In her discussion of the "paradigm wars," for example, Datta (1994) concludes:

> [T]he differences [between the qualitative and quantitative paradigms] are less sharp in practice than in theoretical statements. The best examples of both paradigms seem actually to be mixed models. . . . Perhaps this is not surprising. . . . Most evaluations are conducted under many constraints. These include relatively short time frames, relatively little money, often intractable measurement challenges. . . . In most circumstances, evaluators have to do the best they can and need more, not fewer, approaches on which they can draw. (p. 67)

Some qualitative research crosses the bridge between qualitative and quantitative methods. If we survey clients of a program and ask them to tell us, in their own words, about their experiences with the program, we end up with narrative. Likewise, if we interview persons who delivered the program, questioning them in depth about their perceptions of the program environment, including the clients they serve, we again have narrative. To use this in the evaluation, we need to sort it, organize it, and interpret it. To categorize it we may conduct a **thematic analysis**, looking for groups of similar word or statement clusters.

If we look at the qualitative findings from the client survey, our thematic analysis would give us categories of different themes, and could give us the numbers of times we detected client responses that fit each theme. If we had done our work well, the themes we discerned would cover the range of issues raised by clients in their responses to the survey question and tell us how often each issue was raised.

But we have also created a *nominal variable,* that is, a variable that has mutually exclusive and jointly exhaustive categories. Having done so, we could report the frequencies and relative percentages of themes. We could even cross-classify the frequencies of themes with other nominal variables (the gender of our clients, for example). If we cross-classified themes by other variables, we might even go so far as to test for the statistical significance of the associations: for example, was the distribution of themes significantly different for men than it was for women?

Nominal variables can, in some situations, be just as amenable to statistical manipulations as "higher" levels of measurement—the statistical tools are different, but we are still adding, subtracting, and counting as we do the analysis.

Key Differences Between Qualitative and Quantitative Evaluation Approaches

Table 5.1 is a listing of some of the differences that have been cited between qualitative and quantitative evaluation approaches. The two lists are intended to convey principles and practices that evaluators might use to distinguish the two approaches. It is worth noting that the differences in Table 5.1 are not absolute. Because our views of the roles of qualitative and quantitative evaluation continue to change, it is possible to find advocates for and examples of the view that qualitative data can be the main source of information in randomized experiments (Miles & Huberman, 1994).

Table 5.1 suggests an image of qualitative program evaluation that, although it does not convey the differences among qualitative approaches, does highlight some common central features.

Table 5.1 Key Differences Between Qualitative and Quantitative Evaluation Approaches

Qualitative Evaluation Is Often Characterized by	*Quantitative Evaluation Is Often Characterized by*
• Inductive approach to data gathering, interpretation, and reporting • Holistic approach: finding **gestalts** for the evaluation results • **Verstehen:** understanding the subjective lived experiences of program stakeholders (discovering their truths) • Using natural language throughout the evaluation process • In-depth, detailed data collection • Use of case studies • The evaluator as the primary measuring instrument • A **naturalistic approach:** does not explicitly manipulate the setting	• Research hypotheses and questions that are tested in the evaluation • Finding patterns that either corroborate or disconfirm particular hypotheses and answer the evaluation questions • Understanding how social reality, as observed by the evaluator, corroborates or disconfirms hypotheses and evaluation questions • Emphasis on measurement procedures that lend themselves to numerical representations of variables • **Representative samples** of stakeholder groups • Use sample sizes with sufficient statistical power to detect expected outcomes • Measuring instruments that are constructed with a view to making them reliable and valid • Evaluator control and ability to manipulate the setting, which improves the internal validity, the statistical conclusions validity, and the construct validity of the research designs

In qualitative evaluation, emphasis is placed on the uniqueness of human experiences, eschewing efforts to impose categories or structures on experiences, at least until they are fully rendered in their own terms.

Qualitative program evaluation tends to build from these experiences upwards, seeking patterns but keeping an open stance toward the new or unexpected. The **inductive approach** starts with “the data,” namely, narratives, direct and indirect (unobtrusive) observations, interactions between stakeholders and the evaluator, documentary evidence, and other sources of information, and then *constructs* an understanding of the program. Putting together the themes in the data, weighting them, verifying them with stakeholders, and finally, preparing a document that reports the findings and conclusions is part of a **holistic approach** to program evaluation. A holistic approach also, like an empowerment approach, entails taking into account and reporting different points of view on the program, its operations, and its effects on stakeholders. Thus, an evaluation is not just conducted from the program manager's standpoint, but takes into account clients' viewpoints, as well as other stakeholders' views. Later in this chapter we will provide further suggestions for structuring a qualitative evaluation research project. Many of the major steps in implementing a qualitative evaluation design are, however, fundamentally parallel to the steps in a quantitative evaluation design:

1. Data collection
2. Analysis of the data
3. Writing of the report
4. Dissemination of the report
5. Making changes, based on the evaluation

Similarly, the major questions to be addressed in designing the evaluation have parallels with those in a quantitative evaluation approach.

Qualitative evaluations tend to be **naturalistic**, that is, they do not attempt to control or manipulate the program setting. Instead, the evaluator works with the program as it is and works with stakeholders as they interact with or perform their duties in relation to the program or with each other. Naturalistic also means that natural language is used by the evaluator—the same words that are used by program stakeholders. There is no separate "languages of research design," for example, and usually no separate language of statistics.

In qualitative evaluations, the evaluators themselves are the principal measuring instrument. There is no privileged perspective in an evaluation. It is not possible for an evaluator to claim objectivity. Evaluator observations, interactions, and renderings of narratives and other sources of information are a critical part of constructing patterns, and creating an evaluation report. A principal means of gathering data is face-to-face interviews/conversations. Mastering the capacity to conduct interviews and observations while recording the details of such experiences is a key skill for qualitative program evaluators.

In contrast, quantitative evaluation tends to emphasize hypotheses (embedded in the program logic, for example) or evaluation questions, which generally reflect a limited number of possible stakeholder perspectives. Typically, a key evaluation question is whether the program produced/caused the observed outcomes, that is, *was the program effective?.*

Table 5.2 Summary of Key Questions in Conducting Qualitative Evaluation Assessments and Evaluation Studies

1. Who are the client(s) for the evaluation?
2. What are the questions and issues driving the evaluation?
3. What resources are available to do the evaluation?
4. What has been done previously?
5. What is the program all about?
6. What kind of environment does the program operate in and how does that affect the comparisons available to an evaluator?
7. Which research design alternatives are desirable and appropriate?
8. What information sources are available/appropriate, given the evaluation issues, the program structure and the environment in which the program operates?
9. Given all the issues raised in points 1 to 8, which evaluation strategy is least problematical?
10. Should the program evaluation be undertaken?

Quantitative evaluation is concerned with validity and, in particular, threats to internal validity that would undermine efforts to assess the incremental outcomes of the program. Concerns with internal validity and statistical conclusions validity, in particular, usually mean that quantitative evaluators prefer having some control over the program design, program implementation, and the evaluation processes. Naturalistic settings, however, present limited opportunities in terms of controlled research design options, and render efforts to attribute causal relationships problematical, at least from an internal validity perspective.

Mohr (1999) agrees that most evaluators have tended to question the value of qualitative methods for determining causality in evaluations. The basic problem is that our conventional notion of causality, discussed in Chapter 3, requires some kind of comparison to see what would have happened without the program. In other words, seeing whether the program caused the actual observed outcomes involves establishing what the pattern of outcomes would have been *without* the program. The logic of this process is that if X (the program) caused Y (the observed outcome), then both X and Y occurred, and if X had not occurred, then neither would Y have occurred. If there are no rival hypotheses to cause Y to occur in the absence of X, the counterfactual condition can be demonstrated, and we can conclude that X caused Y.

Mohr suggests that we invoke an alternative model of causality—the modus operandi approach introduced by Scriven (1976). This alternative model is demanding—it requires that there be "physical" connection between the causal variable (X) and Y. The example Mohr uses is from medicine: it is possible to work backwards from a set of symptoms to determine whether a patient had a recent heart attack. A blood test will show whether there are enzymes present that are uniquely associated with a heart attack. Where it is possible to connect the cause physically to the effect, it is not necessary to have a counterfactual comparison—a single case permits us to determine causality.

In principle then, even single cases, which typify some qualitative evaluations, can contribute to understanding cause and effect linkages. Although establishing physical causality in program evaluations can be daunting, Mohr notes that the weakest research designs get used in evaluations often, and we seem to be able to learn by using them:

> Furthermore, ex-post facto or observational studies are well known to be pathetic in terms of internal validity, yet they are in constant use, probably because there is often so much to be learned from them nevertheless. (Mohr, 1999, p. 80)

Parenthetically, it is worth noting that among quantitative evaluators there has been considerable debate over the importance of internal validity. Cook (1991), for example, has argued that internal validity has to be a central concern of evaluations because assessing whether the program really did cause the observed outcomes is a key part of knowing whether the program worked and, thus, whether to ask how the findings and conclusions might be generalized.

Lee Cronbach, on the other hand, has argued that although internal validity should not be ignored, the key issue is external validity: the issue being the generalizability of the evaluation results (Cronbach in Cook, 1991). Cronbach looked at this issue from a practitioner's standpoint and concluded that pinning down causal linkages was so difficult that it tended to absorb far too much of an evaluator's efforts, when the real issue was what could be generalized from a given evaluation and used elsewhere. In their recent book, Shadish, Cook, and Campbell (2002) have come some ways toward Cronbach's position. No longer is internal validity the arbiter of the value of an evaluation. External validity is more prominent now in their schema of four kinds of validities, and they agree with Cronbach that external validity can be viewed independently of internal validity in a program evaluation.

Later in this chapter we will examine the sets of validity issues applicable to qualitative evaluation research, and compare them to the four kinds of validity discussed in Shadish, Cook, and Campbell (2002).

STRUCTURING QUALITATIVE PROGRAM EVALUATIONS

The issue of how much structure to impose on a qualitative evaluation is contentious. At one end of the spectrum, some evaluators advocate an unstructured approach, which does not depend on stakeholders articulating their evaluation questions and expectations for the evaluation process in advance. Evaluators in such settings explore stakeholder viewpoints, and as information is gathered, inductively construct issues that are supported by evidence. These issues, in turn, can be used to guide further data collection. The goal is a full, authentic representation of issues and views that stakeholders have contributed.

The other end of the spectrum is perhaps more common. Program evaluators can construct **conceptual frameworks,** which then guide the evaluation, including what data to collect, who to interview, and what to ask them.

One way to look at the issue of structure is in terms of more specific topics:

1. Identifying evaluation questions and issues
2. Identifying research designs and comparisons
3. Sampling methods
4. Data collection instruments
5. Collecting and coding qualitative data

In expanding these steps, we provide some examples, including a qualitative research study related to home nursing visitation programs. In that study (Byrd, 1999) describes an 8-month field study of 53 home visits to at-risk infants by one nurse. Another study, by McNaughton (2000), provides a synthesis of seventeen qualitative studies of nurse home visitation. These studies relate to the experimental research conducted by Olds and his colleagues in the United States (see Chapter 3). Olds' work has emphasized understanding whether home visits by nurses improve the well-being of children and mothers, but does not describe the actual processes that nurses use in developing relationships with their clients.

Identifying Evaluation Questions and Issues in Advance

In all program evaluation situations, time and money are limited. Usually, evaluations are motivated by issues or concerns raised by program managers or other stakeholders. Those issues constitute a beginning agenda for the evaluation process. The evaluator will usually have an important role in defining the evaluation issues and may well be able to table additional issues. But it is quite rare for an evaluation client or clients to support a fully exploratory evaluation.

In the Byrd (1999) study, the researcher explained that "[t]he processes public health nurses use to effectively work with . . . families are not adequately described" and that "[d]escribing and interpreting the process of home visiting can contribute to the development and refinement of theory that provides a meaningful framework for practice and for studies examining the efficacy of these challenging home visits" (p. 27). The key issue was that although many quantitative studies had done comparisons between program groups and control groups in terms of nurse home visitations, there had been no in-depth studies that looked at what typically occurs during nurse home visits. The study, then, set out to address that gap.

The McNaughton (2000) study was conducted later and looked at the by-then larger selection of qualitative studies on the effects of nurse home visitations. The issue was that before the qualitative studies had been done, "it was difficult to determine aspects of nursing interventions that were or were not effective" (p. 405). The goal of this qualitative research was to gather and analyze qualitative nurse home visitation studies "to provide an organized and rich description of public health nursing practice based on identification of common elements and differences between research reports" (p. 405.)

Identifying Research Designs and Appropriate Comparisons

Qualitative data collection methods can be used in a wide range of research designs. Although they can require a lot of resources, qualitative methods can be used even in fully randomized experiments, where the data are compared and analyzed with the goal of drawing conclusions around the program's incremental outcomes.

More typically, the comparisons are not structured around experimental or even quasi-experimental research designs. Instead, **implicit designs** are often used. The emphasis, then, is on what kinds of comparisons to include in the data collection and analysis.

Miles and Huberman (1994) indicate that there are two broad types of comparisons, given that you have collected qualitative data. One is to focus on single cases and conduct analyses on a case-by-case basis. Think of a case as encompassing a number of possibilities. In an evaluation of the Perry Preschool experiment (see Chapter 3), the individual children in the study (program and control groups) were the cases. In an evaluation of a Neighbourhood Integrated Service Teams (NIST) Program in Vancouver, Canada, each NIST was a case; there were a total of 15 in the city and all were included in the evaluation (Talarico, 1999). In the Byrd (1999) study of home nursing, the case was the one nurse who was observed for 8 months, and the researcher aggregated the results from observing 53 home visits, in order to produce "a beginning typology of maternal-child home visits" (p. 31).

Cases are, in the parlance of Chapter 4, ***units of analysis***. When we select cases in a qualitative evaluation, we are selecting units of analysis.

Cases can be described in depth. Events can be reconstructed as a chronology. This is often a very effective way of describing a client's interactions with a program, for example. Cases can include quantitative data. In the NIST evaluation, it was possible to track and count the levels of activities for each NIST from the program's inception in 1996 to the evaluation in 1999.

The second kind of comparison using cases is *across* cases. Selected program participants in the Perry Preschool experiment, for example, were compared using qualitative analysis. Each person's story was told, but their experiences were aggregated: men versus women, for example. In the McNaughton (2000) study, the research reports were analyzed individually first ("within case" analysis), and then later the author conducted cross-case analysis that "consisted of noting commonalities and differences between the studies" (p. 407).

Cases can be compared across program sites. If a program has been implemented in a number of geographic locations, it might be important to conduct **case studies** of clients (for example) in each area, and then compare client experiences across areas. There is no reason why such qualitative comparisons could not also be complemented by quantitative comparisons. Program sites might be compared over time on the levels of program activities (outputs) and client satisfaction, and provide perceptions of program outcomes.

Identifying Appropriate Samples

Qualitative **sampling strategies** generally deliberately select cases. Contrast this approach with a quantitative evaluation design that emphasizes random samples of cases. Typically, the total number of cases sampled is quite limited, so the selection of cases becomes critical. Note that the Byrd (1999) study followed just *one* nurse over an 8-month period. The author notes "[i]nitially, observations with several nurses were planned, but well into the fieldwork, it became evident that prolonged full engagement with just one nurse was critical" (p. 28). Initially "the researcher observed all home visits scheduled during the fieldwork day" but as the patterns of the process emerged, "the emphasis shifted to describing and elaborating the patterns, so the research asked to accompany the nurse on visits anticipated to follow a specific pattern" (p. 28).

Table 5.3 is a typology of sampling strategies developed by qualitative researchers. This version of the table is adapted from Miles and Huberman (1994, p. 28). The 16 types of sampling strategies summarized in Table 5.3 are similar to the 14 purposeful sampling strategies summarized in Patton's (2003) Qualitative Evaluation Checklist.

Among the strategies identified in Table 5.3, several tend to be used more frequently than others. **Snowball sampling** relies on a chain of informants, who are themselves contacted, perhaps interviewed, and asked who else they can recommend, given the issues being canvassed. Although this sampling procedure is not random and may not be representative, it usually yields informed participants. One rough rule of thumb to ascertain when a snowball sample is "large enough" is to note when themes and issues begin to repeat themselves across informants.

The Byrd (1999) study would probably be considered an example of a case of *intensity* sampling, one that "manifests the phenomenon intensely, but not extremely." The author makes the point in her article that the nurse may not have been typical, so while the case was an in-depth one, it could not be assumed that the typology would apply broadly to all, or even most, nurses conducting home visits.

Sampling *politically important* cases is often a component of qualitative sampling strategies. In a qualitative study of stakeholder viewpoints in an intergovernmental economic development agreement, the 1991–1996 Canada/Yukon Economic Development Agreement (McDavid, 1996), the evaluator initially relied on a list of suggested interviewees, which included public leaders, prominent business owners, and the heads of several interest group organizations (the executive director of the Yukon Mining Association, for example). Interviews with those persons yielded additional names of persons who could be contacted, some of whom were interviewed, and others who were willing to suggest further names (McDavid, 1996).

The McNaughton (2000) study would be considered an example of **criterion sampling**, and their selection from the qualitative nursing studies was partly described as follows:

> Studies included in the analysis were written in English and were published articles or doctoral dissertations reporting original research. Only research investigating home visits between PHNs [public health nurses] and mothers of young children was included. In addition, only reports

Table 5.3 Sampling Strategies for Qualitative Evaluations

Type of Sampling	*Purpose*
Maximum variation	Documents variation and identifies important common patterns
Homogeneous	Focuses, reduces, simplifies, facilitates group interviewing
Critical case	Permits logical generalization and maximum application of information to other cases
Theory based	Finding examples of a theoretical construct and thereby elaborate and examine it
Confirming and disconfirming cases	Elaborating initial analysis, seeking exceptions, looking for variation
Snowball or chain	Identifies cases of interest from people who know people who know what cases are information-rich
Extreme or deviant case	Learning from highly unusual manifestations of the phenomenon of interest
Typical case	Highlights what is normal or average
Intensity	Information-rich cases that manifest the phenomenon intensely, but not extremely
Politically important cases	Attracts desired attention or avoids attracting undesired attention
Random purposeful	Adds credibility to sample when potential purposeful sample is too large
Stratified purposeful	Illustrates subgroups; facilitates comparisons
Criterion	All cases that meet some criterion; useful for quality assurance
Opportunistic	Following new leads; taking advantage of the unexpected
Combination or mixed	Triangulation, flexibility, meets multiple interests and needs
Convenience	Saves time, money, and effort, but at the expense of information and credibility

> using qualitative design and published after 1980 were reviewed. Seventeen studies were retrieved and 14 included in the final analysis. (p. 406)

Opportunistic sampling takes advantage of the inductive strategy that is often at the heart of qualitative interviewing. An evaluation may start out with a sampling plan in mind (picking cases that are representative of key groups or interests) but as interviews are completed, a new issue may emerge that needs to be explored more fully. Interviews with persons connected to that issue may need to be conducted.

Mixed sampling strategies are common. As was indicated for the Canada/Yukon Economic Development Agreement project, an initial sample that was dominated by politically important persons was combined with a snowball sample. In pursuing mixed strategies, it is valuable to be able to document how sampling decisions were made. One of the criticisms of qualitative samples is that they have no visible rationale—they are said to be drawn capriciously and the findings cannot be trusted. Even if sampling techniques do not include random or stratified selection methods, documentation can blunt criticisms that target an apparent lack of a sampling rationale.

Structuring Data Collection Instruments

Typically, qualitative data collection components of program evaluations are structured to some extent. It is very unusual to conduct interviews, for example, without at least a general agenda of topics. Additional topics can emerge, and the interviewer may wish to explore connections among issues that were not anticipated in the interview plan. But the reality in program evaluations is that resource constraints will mean that interviews are focused and at least semi-structured.

Qualitative survey instruments generally use **open-ended questions**, unlike highly structured quantitative surveys, which typically have a preponderance of **closed-ended questions.** Table 5.4 is a summary of the open-ended questions that were included in all interviews conducted for the Canada/Yukon Economic Development Agreement project (McDavid, 1996).

Each interview took at least 1 hour and most lasted 2 or more hours. The open-ended questions in Table 5.4 were structured to follow the evaluation questions that the overall program evaluation was expected to answer. The stakeholder interviews were intended to provide an independent source of information on the evaluation questions, and the findings from the interviews were integrated into the overall evaluation report (McDavid, 1996).

Each interview was tape-recorded, and the researcher conducting the interviews used the tapes to review and fill in his interview notes for each interview. Because the questions were organized around key issues in the overall evaluation, the analysis of the interview data focused on themes within

Table 5.4 Open-Ended Questions for the Economic Development Agreement Stakeholders Project

Canada/Yukon Economic Development Agreement Evaluation Stakeholder Interview Questions (Note date, time, location, name/position of interviewee)

Begin by introducing interviewer, reviewing purpose of interview, and requesting permission to tape-record interview. Note: Only the interviewer will have access to the tape of the interview.

1. What is your interest in and/or involvement in the 1991–1996 Economic Development Agreement between the Government of Canada and theYukon Territorial Government?
   - Which of the six Cooperation Agreements are you involved with?
   - Were you involved in the previous EDA (1985–1989)?
   - How?

2. How much (if any) do you know about the EDA and the six Cooperation Agreements?
   - Do you have a good idea of how the whole EDA works?
   - Which of the CAs are you familiar with?
   - (If appropriate) How did you find out about the EDA?

3. How would you assess the strengths and weaknesses of the organization and the administration of the EDA and the CAs?
   - Particular strong points?
   - Particular weak points?
   - Suggestions for improving the organization/administration?
   - Barriers to improving the organization/administration?

4. Can you think of any EDA-funded projects that you would consider to be successful? Which projects, if any, were/are they? Why were they successful? Are there specific reasons for their success?

5. Which EDA projects, if any, would you consider to be failures, that is, they did not achieve any of their objectives? Are there specific reasons for their failure?

6. Now, understanding that we are asking for your overall assessment, how effective has each CA been in meeting its objectives (refer to list of Cooperation Agreement objectives, if necessary)?
   - Which CA(s) has/have been the most successful? Why?
   - Which CA(s) has/have been the least successful? Why?

   (Note: Some interviewees may not be familiar with all CAs.)

7. How effective would you say the EDA as a whole has been in meeting its overall objectives (refer to them in the list, if necessary)?
   - What part(s) of the EDA has/have been most successful in achieving these objectives?
   - What part(s) has/have been least successful?
   - Why?

(Continued)

Table 5.4 (Continued)

8. Have there been any unintended impacts from EDA-funded activities? Can you suggest some examples?

9. Has the EDA been equitably accessible to all Yukoners? Which CAs have been relatively more accessible? Which CAs have been less accessible? (Note: If necessary, remind interviewees that the four Federally-mandated target groups (Aboriginal peoples, women, handicapped persons, and visible minorities) are included in this question.)

10. If you were in a position to offer your advice on re-designing the EDA, which CAs would you change to make them more effective (more likely to achieve their objectives)?

    - What changes would you make?
    - What particular parts (elements) would you add or delete?

    (Note: Some interviewees may not be able to answer this question, given their knowledge of the EDA.)

11. Thinking of economic development as a broad issue for the Yukon, are there other kinds of programs, besides an EDA, that might be more cost-effective in achieving the objectives of the current EDA (refer to list of objectives again, if necessary)?
    - What would this/these program(s) look like?

12. Any other comments?

Thank you for your time and input into the evaluation of the EDA.

SOURCE: McDavid, 1996.

each of the issue areas. In some interviews, information relevant to one question was offered in response to another and was reclassified accordingly.

Structuring data collection instruments has several limitations. By setting out an agenda, the qualitative evaluator may miss opportunities to follow an interviewee's direction. If qualitative evaluation is, in part, about reconstructing others' lived experiences, structured instruments, which imply a particular point of view on what is important, can largely limit opportunities to empathetically understand stakeholders' viewpoints.

It may be appropriate in an evaluation to begin qualitative data collection without a fixed agenda, to learn what the issues, concerns, and problems are so that an agenda can be established. In the Byrd (1999) observational study of home nurse visitations, the data collection description provides an interesting example of a relatively fluid project, as the collection of data was guided as the analysis was occurring:

> Simultaneous data collection, field note recording, and analysis were focused on the nurse's intentions, actions, and meanings as she anticipated, enacted, or reflected on her visits. In an effort to understand her thinking as she did home visiting, this nurse was informally interviewed before and after home visits. Interviews included probing questions to promote self-reflection. The investigator explored the nurse's perceptions of the reasons for the home visit; concerns before, during, and after the home visit; and her perceptions of what she was trying to accomplish, as well as the anticipated consequences of her actions. (p. 28)

A practical limitation on the use of unstructured approaches is their cost. Often, evaluation budgets do not permit us to solely conduct unstructured interviews, and then consume the resources needed to organize and present the findings.

Collecting and Coding Qualitative Data

A principal means of collecting qualitative data is interviews. Although other ways are also used in program evaluations (e.g., documentary reviews/analyses, open-ended questions in surveys, direct observations), face-to-face interviews are a key part of qualitative data collection options.

Table 5.5 summarizes some important points to keep in mind when conducting face-to-face interviews. The points in Table 5.5 are not exhaustive, but are based on the writers' experiences of participating in qualitative interviews and qualitative evaluation projects. Patton (2003) includes sections in his Qualitative Evaluation Checklist that focus on field work and open-ended interviewing. Patton's experience makes his checklists a valuable source of information for persons involved in qualitative evaluations.

Table 5.6 offers some helpful hints about analyzing qualitative data, again, principally from face-to-face interviews. As Patton (2003) reiterates in the Qualitative Evaluation Checklist, it is important that the data are effectively analyzed "so that the qualitative findings are clear, credible, and address the relevant and priority evaluation questions and issues" (p. 10).

Coding of the data gathered in the Byrd (1999) study was described as follows:

> initial analysis focused on fully describing the process of home visiting. Later, coding and theoretic memos—both analytic techniques from grounded theory—were used. Coding is assigning conceptual labels to

Table 5.5 Some Basics of Face-to-Face Interviewing

General Points to Keep in Mind

- Project confidence and be relaxed—you are the measuring instrument, so your demeanor will affect the entire interview.
- Inform participants—make sure they understand why they are being interviewed, what will happen to the information they provide, and that they can end the interview or not respond to specific questions as they see fit (informed consent).
- Flexibility is essential—it is quite possible that issues will come up "out of order" or that some will be unexpected.
- Listening (and observing) are key skills—watch for word meanings or uses that suggest they differ from your understanding. Watch for non-verbal cues that suggest follow-up questions or more specific probes.
- Ask for clarifications—do not assume you know or that you can sort something out later.

Conducting the Interview

- Ask questions or raise issues in a conversational way
- Show you are interested, but non-judgmental
- Look at the person when asking questions or seeking clarifications
- Consider the cultural appropriateness of eye contact
- Pace the interview so that it flows smoothly
- Note taking is *hard work:* the challenge is to take notes, listen, and keep the conversation moving
  - Consider having one researcher conduct the interview while another takes notes
  - Note key phrases
  - If you can, use a tape-recorder
    - Issues of confidentiality are key
    - If you are trying to record sentences verbatim to use as quotes, you may need to stop the conversation for a moment to write
- Your recall of a conversation decays quickly so *take time* to review your notes, fill in gaps, and generally make sense out of what you wrote, *before* you conduct the next interview.
- Pay attention to the context of the interview—are there situational factors (location of the interview, interruptions or interactions with other people) that need to be noted to provide background information as qualitative results are interpreted?

Table 5.6 Helpful Hints as You Analyze Qualitative Data

Getting Started

- Why did you conduct the interviews? How do the interviews fit into the program evaluation?
- What evaluation issues were you hoping could be addressed by the interview data?
- Can the relevant evaluation issues be organized or grouped to help you sort narrative and notes into themes and sub-themes? Can your interview data be categorized by evaluation issue?
- Always do your work in pencil so you can revise what you have done.

Analyzing the Data

- If you have tape-recorded the interviews, you should listen to the tapes as you review your interview notes to fill in or clarify what was said. Some analysts advocate transcribing tapes and using the transcripts as your raw data. That takes a lot of time, and in many evaluations is not practical. It does, however, ensure the accuracy and completeness of the data that you will be analyzing. Keep in mind that even verbatim transcriptions do not convey all the information that was communicated as part of interviews.
- If you have not taped the interviews, read all your interview notes, jotting down ideas for possible themes as penciled marginal notes.
- Pay attention to the *actual words* people have used—do not put words in interviewees' mouths.
- There is a balance between looking for themes and categories and imposing your own expectations. When in doubt, look for *evidence* from the interviews.
- Thematic analysis can be focused on identifying words or phrases that summarize ideas conveyed in interviews. For example, interviews with government program evaluators to determine how they acquired their training, identified themes like: university courses; short seminars; job experience; and other training.
- Thematic analysis can be focused on identifying statements (subject/verb) in a narrative. It may be necessary to set up a database that translates a narrative into a set of equivalent statements that can be analyzed.

Re-read the interviews. Which of the preliminary themes still make sense? Which ones are wrong? What new themes emerge?

- What are the predominant themes? Think of themes as ideas: they can be broad (in which case lots of different sub-themes would be nested within each theme) or they can be narrow, meaning that there will be lots of them.

(Continued)

Table 5.6 (Continued)

- Are your themes different from each other (they should be)?
- Have you captured all the variation in the interviews with the themes you have constructed?
- How will you organize your themes: alternatives might be by evaluation issue/question; or by affect, that is, positive, mixed, negative views of the issue at hand?
- List the themes and sub-themes you believe are in the interviews. Give at least two examples from the interviews to provide a working definition of each theme or sub-theme.
- Read the interviews again, and this time try to fit the text/responses into your thematic categories.
- If there are anomalies, adjust your categories to take them into account.
- There is almost always an "other" category. It should be no more than 10% of your responses/coded information.
- Could another person use your categories and code the text/responses the way you have? Try it for a sample of the data you have analyzed.
- Calculate the percentage of agreements out of the number of categorizations attempted. This is a measure of inter-coder reliability.
- Are there direct quotes that are appropriate illustrations of key themes?

> incidents or events. Memos link observations and help investigators make inferences from the data. For each visit, data were coded as descriptors of the phases of the process, consequences of the process, or factors influencing the process. Similarities and differences in the process, potential consequences, and influencing factors across visits were then compared. Distinct patterns of home visiting emerged from this analysis. (p. 28)

To illustrate, the thematic coding and analysis of "the phases of the process" in this study resulted in the following model of the nurse home visitation process (p. 28):

Prioritizing the visit → Bartering to schedule the visit → Approaching the home and the visit → Entering the home → Gathering the initial information → Making the caregiving judgment → Ending the visit → Feeling haunted and telling a concerned person

THE CREDIBILITY AND GENERALIZABILITY OF QUALITATIVE FINDINGS

Debates about sampling and the ability to generalize from one's data are important—they are at the core of a frequent criticism of qualitative methods. Qualitative evaluators emphasize the value of in-depth, case-based approaches to learning about a program. Given that each stakeholder will offer his or her own world view, it is essential to "peel away" the layers of the onion to gain a full and authentic rendering of a person's experiences, views, and assessments of the program. As Kushner (2000) argues, since "the worth of a program is . . . subject to situational interpretation and contested meaning," program evaluators should seek to "document the lives and work of people and to use that as context within which to 'read' the significance and meaning of programs" (pp. xiv, 11).

This takes time and considerable effort. The main source of data is narrative, and analyzing these data is also time-consuming. In sum, there is a tradeoff between depth (and increasing the validity of the data) and breadth (increasing the representativeness of the data). Qualitative methods focus on fewer cases, but the quality and completeness of the information is viewed by proponents as outweighing any disadvantages due to lack of representativeness.

A challenge for evaluators who use qualitative methods is to establish the credibility and generalizability of their findings, that is, their believability and hence usefulness for stakeholders. Relying on analyses of cases can produce rich, detailed information, but if we cannot address possible concerns about the representativeness of the findings, or the methods used to produce them, our work has not been productive.

Maxwell (2002), in a synthesis of various approaches to validity in qualitative research, outlines five types of understanding and validity that typify qualitative research. His efforts were stimulated by the following observation:

> Proponents of quantitative and experimental approaches have frequently criticized the absence of "standard" means of assuring validity, such as quantitative measurement, explicit controls for various validity threats, and the formal testing of prior hypotheses. (p. 37)

Maxwell outlines a typology that exemplifies the ways that qualitative researchers conceptualize validity. The typology (Table 5.7) includes: descriptive validity, interpretive validity, theoretical validity, generalizability, and evaluative validity. Maxwell's epistemological stance is critical realism, that is, he believes that there is a reality external to our perceptual knowledge of it, but

Table 5.7 Comparing Qualitative Validity with Experimental and Quasi-Experimental Validity

Types of Validity in Qualitative Research[a]	*Definitions of Qualitative Validities*	*Related to the Following Types of Validity in Experimental/ Quasi-Experimental Research*[b]
Descriptive Validity	The factual accuracy of the account (consensus of researchers—intersubjective agreement on the existence of physical and behavioral events); can include descriptive statistics (e.g., frequencies)	Statistical conclusions validity (specifically the reliability of measures)
Interpretive Validity	The meanings of actions or behaviors from participants' perspectives.	No correspondences with validity categories in the Shadish, Cook, and Campbell typology
Theoretical Validity	Focus is on the researcher's constructs—both individual constructs and causal relationships among constructs	Construct validity (how well do the factual accounts link with researcher constructs that interpret them, and how well do factual patterns correspond to relationships among constructs?)
Generalizability • Internal • External	Generalizing to other persons, organizations or institutions within the community Generalizing to other communities, groups or organizations	• Statistical conclusions validity (inferential statistics) • External validity (do the causal relationships hold for variations in persons, treatments, settings and outcomes?)
Evaluative Validity	Judging the appropriateness of actions or events from a values perspective	No correspondence with validity categories in Shadish, Cook, and Campbell

a. From Maxwell, 2002.

b. From Shadish, Cook, and Campbell, 2002.

we cannot know that reality directly. His validity categories are not intended to be a filter to assess or judge the quality of a study. Unlike positivists or post-positivists who use validity categories to discriminate among methods for doing research (randomized control trials are generally superior from an internal validity perspective, for example), Maxwell sees types of validity as fallible and not proscribing particular methodologies. The relevance of validities depends of the circumstances of a given research study.

Table 5.8 Ways of Testing and Confirming Qualitative Findings

1. **Check the cases for representativeness** by comparing case characteristics to characteristics of people (units of analysis) in the population from which the cases were selected.
2. **Check for researcher effects** by asking whether and how the evaluator could have biased the data collection or how the setting could have biased the researcher.
3. **Triangulate data sources** by comparing qualitative findings with other sources of data in the evaluation.
4. **Weigh the evidence** by asking whether some sources of data are more credible than others.
5. **Check outliers** by asking whether "deviant" cases are really that way or, alternatively, the "sample" is biased and the outliers are more typical.
6. **Use extreme cases** to calibrate your findings, that is, assess how well and where your cases sit in relation to each other.
7. **Follow up surprises,** that is, seek explanations for findings that do not fit the overall patterns.
8. **Look for negative evidence,** that is, findings that do not support your own conclusions.
9. **Formulate If/Then statements** based on your findings to see if interpretations of findings are internally consistent.
10. **Look for intervening variables** that could explain key findings—if you have information on these variables, can you rule their influences out, based on your findings?
11. **Replicate findings** from one setting to another one that should be comparable.
12. **Check out rival explanations** using your own data, your judgment, and the expertise of those who know the area you have evaluated.
13. **Get feedback from informants** by summarizing what they have contributed and asking them for their **concurrence** with your summary.

SOURCE: Miles and Huberman (1994, pp. 263–277).

Some, though not all, of the types have commonalities with the types of validity described by Shadish, Cook, and Campbell (2002). But Maxwell stresses that even where there are commonalities, it is important to keep in mind that they do not indicate a shared epistemology with positivist and post-positivist researchers.

Miles and Huberman (1994) have identified 13 separate ways that qualitative data and findings can be queried to increase their robustness. Table 5.8 lists, adapts, and summarizes these checks, together with a brief explanation of what each means.

Although these 13 points all offer complementary ways to increase our confidence in qualitative findings, some are more practical than others. In program evaluations, two of these are more useful:

- Triangulating data sources
- Getting feedback from informants

Feedback from informants goes a long way toward establishing the validity of qualitative data. It does not tell you how representative your cases are, but it does tell you whether you have rendered the information so that it accords with the views of those providing it—that is key to authentically representing their world views.

Triangulation of data sources is important to establish whether findings from qualitative analyses accord with those from other data sources. Typically, complementary findings suggest that the qualitative data are telling the same story as are other data. If findings diverge, then it is appropriate to explore other possible problems: representativeness of the cases, researcher bias, and weighing the evidence are reasonable places to begin.

CONNECTING QUALITATIVE EVALUATION METHODS TO PERFORMANCE MEASUREMENT

Performance measurement has tended to rely on quantitative measures for program constructs. Program or organizational objectives are stated, annual performance targets are established in many performance measurement systems, and the data that are gathered are numerical. Numbers lend themselves to visual displays (graphs, charts) and are relatively easy to interpret (trends, levels). But, for some government agencies and nonprofit organizations, the requirement that their performance be represented in numbers forces the use of measures that are not seen by agency managers to reflect the key outcomes. Nonprofit organizations that mark their progress with clients by seeing individual lives being changed often do not feel that numerical performance measures weigh or even capture these outcomes.

Sigsgaard (2002) has summarized an approach to performance measurement that is called the Most Significant Change (MSC) approach. Originally designed for projects in developing nations, where aid agencies were seeking an alternative to numerical performance measures, the MSC approach applies qualitative methods to assessing performance. It has something in common with the Shoestring Evaluation approach (Bamberger, Rugh, Church, & Fort, 2004)—both are designed for situations where evaluation resources are very limited, but there is a need to demonstrate results and do so in ways that are defensible.

Sigsgaard (2002) describes how a Danish international aid agency (Mellemfolkeligt Samvirke) adopted the MSC approach as an alternative to the traditional construction of quantitative logic models of projects in developing countries. The main problem with the logic modeling approach was the inability of stakeholders to define objectives that were amenable to quantitative measurement.

The MSC approach involves an interviewer or interviewers (who have been briefed on the process and intent of the approach) asking persons who have been involved in the project (recipients/beneficiaries of the project) to identify positive or negative changes they have observed over a fixed time, for one or more domains of interest. Examples of a domain might be health care in a village involved in an aid project, or farming in a rural area where a project had been implemented. By eliciting both positive and negative changes, there is no bias towards project success. Then these same persons are asked to indicate which change is the most significant and why.

By interviewing different stakeholders, a series of change-related stories are recorded. Although they might not all relate to the project or to the project's objectives, they provide authentic views on how participants in the MSC interviews see their world and the project in it.

The performance stories are reviewed by different governance levels (boards) in the donor organization (within and outside the country), and from among them, the most significant stories (ultra-most significant) are selected along with reasons for their choices. Essentially, the set of performance stories are shared and discussed and finally winnowed to a smaller set. Performance stories are verified by additional investigation and are then used to guide any changes that are implied by the results that are communicated via the stories.

Sigsgaard (2002) sums up the experience of his aid organization with the MSC approach to qualitative performance measurement:

> The process of verification, and the curiosity aroused by the powerful data collected, will stimulate the country offices as well as the partners to supplement their knowledge through use of other, maybe more refined and controlled measures. The MSC system is only partially participatory. Domains of interest are centrally decided on, and the sorting of stories according to significance is hierarchic. However, I believe that the use of and respect for peoples' own indicators will lead to participatory methodologies and "measurement" based on *negotiated indicators* where all stakeholders have a say in the actual planning of the

> development process. Some people in the MS [Mellemfolkeligt Samvirke] system have voiced a concern that the MSC method is too simple and "loose" to be accepted by our source donor, Danida, and our staff in the field. The method is not scientific enough, they say. My computer's Thesaurus program tells me that science means knowledge. I surely can recommend the Most Significant Change method as scientific. (p. 11)

● THE POWER OF CASE STUDIES

One of the great appeals of qualitative evaluation is the ability to render experiences in convincing detail. Narrative from even a single case, rendered to convey a person's own words and feelings, is a very powerful way to draw attention to an issue or a point of view.

Most of us respond favorably to stories, to narratives that chronicle the experiences of individuals. In the context of program evaluations, it is often much easier to communicate key findings by using case examples. For many clients, tables do not convey a lot of intuitive meaning. Graphs are better; but narratives, in some cases, are best. Patton (2003), in his checklist for qualitative evaluations, suggests that qualitative methods are best suited for telling stories:

> Qualitative methods are often used in evaluations because they tell the *program's story* by capturing and communicating the *participants' stories.* Evaluation case studies have all the elements of a good story. They tell what happened when, to whom, and with what consequences. (p. 2)

In the mass media, typically news stories focus on individuals, and a single well-stated opinion or carefully presented experience can have important public policy implications. The tragic death of a single child in British Columbia, Canada in 1994 at the hands of his mother became the basis for the Gove Commission (Gove, 1995) and, ultimately, the reorganization of all existing child protection functions into the provincial Ministry for Children and Families in 1996.

In program evaluations, case studies often carry a lot of weight, simply because we can relate to the experiences of individuals more readily than we can understand the aggregated/summarized experiences of many. Even though single cases are not necessarily representative, they are often treated as if they contained *more data* than just one case. For program evaluators, there is both an opportunity and a caution in this. The opportunity is to be able to use cases and qualitative evidence to render evaluation findings more credible and, ultimately, more useful. But the caution is to conduct qualitative evaluations (or the qualitative components of multi-source evaluations) so that they are *methodologically defensible* as well as being persuasive.

SUMMARY ●

Qualitative evaluation methods are an essential part of the range of tools that evaluators call upon in their practice. Since the 1970s, when qualitative evaluation methods were first introduced as an alternative to the then orthodox experimental/quasi-experimental paradigm, the philosophical underpinnings and methodological requirements for sound qualitative evaluation have transformed the evaluation profession. Debates continue about the relative merits of positivistic and constructivist approaches to evaluation, but many evaluators have come to the view that pragmatically, it is desirable to mix qualitative and quantitative methods—they have complementary strengths and the weaknesses of one approach can be mitigated by calling upon the other approach.

Qualitative approaches to evaluation are themselves, diverse. Some proponents share the same epistemological beliefs as do practitioners who rely on quantitative methods—that there is a reality we share and can know (to varying degrees) through our efforts to measure aspects of it. Other qualitative evaluators have embraced one or another phenomenological approach—the underlying assumptions about the way we know do not include the belief that there is one (social) reality we share. Rather, each of us has our own "world" and the challenge for evaluators is to develop methods to learn about each person's world, render what has been learned in ways that are authentic, and find ways of working with those perspectives in the evaluation process.

Qualitative evaluation often relies on case studies—in-depth analyses of individuals (as units of analysis) who are stakeholders in a program. Case studies, rendered as stories are an excellent way to communicate the personal experiences of those connected with a program. We, as human beings, have tended to be storytellers—indeed, stories and songs were the ways we transmitted culture before we had written language. Case studies convey meaning and emotion, rendering program experiences in terms we can all understand.

Although performance measurement has tended to rely on quantitative indicators to convey results, there are alternatives that rely in qualitative methods to elicit performance stories from stakeholders. In settings where data collection capacities are very limited, qualitative methods offer a feasible and effective way to describe and communicate performance results.

● DISCUSSION QUESTIONS

1. What is a paradigm? What does it mean to say that paradigms are incommensurable?
2. What is Patton's pragmatic approach to evaluation?
3. What are the key characteristics of qualitative evaluation methods?
4. What does it mean for an evaluation to be naturalistic?
5. What is snowball sampling?
6. Suppose that you have an opportunity to conduct an evaluation for a state agency that delivers a program for single mothers. The program is intended to assist pregnant women with their first child. The program includes home visits by nurses to the pregnant women and regular visits for the first 2 years of the child's life. The objective of the program is to improve the quality of parenting by the mothers, and hence, improving the health and well-being of the children. The agency director is familiar with the quantitative, experimental evaluations of this kind of program in other states and wants you to design a qualitative evaluation that focuses on what actually happens between mothers and children in the program. What would your qualitative evaluation design look like? What qualitative data collection methods would you use to see what was happening between mothers and children? How would you determine whether the quality of parenting had improved, as a result of the program?

● REFERENCES

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation, 25*(1), 5–37.

Byrd, M. E. (1999). Questioning the quality of maternal caregiving during home visiting. *Image: Journal of Nursing Scholarship, 31*(1), 27–32.

Cook, T. D. (1991). Clarifying the warrant for generalized causal inferences in quasi-experimentation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century. Yearbook of the National Society for the Study of Education* (90th ed., Pt. 2, pp. xiv, 296). Chicago: National Society for the Study of Education; University of Chicago Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally College Publishing.

Datta, L. E. (1994). Paradigm wars: A basis for peaceful coexistence and beyond. *New Directions for Program Evaluation, 61,* 53–71.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Fetterman, D. M., (1996). Empowerment evaluation: An introduction to theory and practice. In D. M. Fetterman, S. Kaftarian, & A. Wandersman (Eds.), *Empowerment evaluation: Knowledge and tools for self-assessment and accountability.* Thousand Oaks, CA: Sage.

Gove, T. J. (1995). *Report of the Gove Inquiry into Child Protection in British Columbia: Executive Summary.* Retrieved August 5, 2004, from http://www.qp.gov.bc.ca/gove/gove.htm

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

Guba, E. G., & Lincoln, Y. S. (2001). *Guidelines and checklist for constructivist (a.k.a. Fourth Generation) evaluation.* Retrieved August 4, 2004, from http://www.wmich.edu/evalctr/checklists/constructivisteval.pdf

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science.* Cambridge, UK: Cambridge University Press.

House, E. R. (1994). Integrating the quantitative and qualitative. *New Directions for Program Evaluation, 61,* 13–22.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kushner, S. (2000). *Personalizing evaluation.* London, Thousand Oaks, CA: Sage.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs.* San Francisco: Jossey-Bass.

Maxwell, J. A. (2002). Understanding and validity in qualitative research. In A. M. Huberman & M. B. Miles (Eds.), *The qualitative researcher's companion* (pp. 37–64). Thousand Oaks, CA: Sage.

McDavid, J. C. (1996). Summary report of the 1991–1996 Canada/Yukon EDA evaluation. Ottawa, ON: Department of Indian and Northern Affairs.

McNaughton, D. B. (2000). A synthesis of qualitative home visiting research. *Public Health Nursing, 17*(6), 405–414.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Mohr, L. B. (1999). The qualitative method of impact analysis. *American Journal of Evaluation, 20*(1), 69–84.

Murphy, E., Dingwall, R., Greatbatch, D., Parker, S., & Watson, P. (1998). Qualitative research methods in health technology assessment: A review of literature. *Health Technology Assessment, 2*(16), 1–274.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2003). *Qualitative evaluation checklist.* Retrieved August 4, 2004, from http://www.wmich.edu/evalctr/checklists/qec.pdf

Reichardt, C. S., & Rallis, S. F. (1994). Qualitative and quantitative inquiries are not incompatible: A call for a new partnership. *New Directions for Program Evaluation, 61,* 85–91.

Scriven, M. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 108–118). Beverly Hills, CA: Sage.

Scriven, M. (In progress, unpublished). *Causation.* New Zealand: University of Auckland.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Sigsgaard, P. (2002). MCS approach: Monitoring without indicators. *Evaluation Journal of Australasia, 2*(1), 8–15.

Talarico, T. (1999). *An evaluation of the Neighbourhood Integrated Service Team program.* Unpublished Master's thesis, University of Victoria, British Columbia, Canada.

The Evaluation Center. (2001). *The evaluation checklist project.* Retrieved July 13, 2004, from http://www.wmich.edu/evalctr/checklists/qec/index.htm