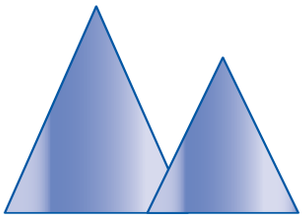
CHAPTER

9 Enhancing the Quality and Credibility of Qualitative Studies



The Medieval alchemical symbol for fire was a single triangle, also the modern symbol for triangulation in geometry, trigonometry, and surveying: the process of locating an unknown point by measuring angles to it from known points. Triangulation in qualitative inquiry involves gathering and analyzing multiple perspectives, using diverse sources of data, and during analysis, using alternative frameworks.

The double-triangle symbol, shown here, represented *strong fire* in alchemy. Strong fire was needed to ensure that the transformative process would work. Building and sustaining a *strong fire* required quality materials, good ventilation, and ongoing monitoring. Using a strong fire required skill, experience, and rigorous implementation of the transformative process to achieve the desired effects. Strong fire produces both intense heat and bright illumination. Alchemists who could properly build, sustain, and appropriately use strong fire were held in high esteem, had great credibility, and produced much-valued products.

Interpreting Truth

A young man traveling through a new country heard that a great Mulla, a Sufi guru with unequaled insight into the mysteries of the world, was also traveling in that region. The young man was determined to become his disciple. He found his way to the wise man and said, "I wish to place my education in your hands that I might learn to interpret what I see as I travel through the world."

After six months of traveling from village to village with the great teacher, the young man was confused and disheartened. He decided to reveal his frustration to the Mulla.

"For six months I have observed the services you provide to the people along our route. In one village you tell the hungry that they must work harder in their fields. In another village you tell the hungry to give up their preoccupation with food. In yet another village you tell the people to pray for a richer harvest. In each village the problem is the same, but always your message is different. I can find no pattern of Truth in your teachings."

The Mulla looked piercingly at the young man.

"Truth? When you came here you did not tell me you wanted to learn Truth. Truth is like the Buddha. When met on the road it should be killed. If there were only one Truth to be applied to all villages, there would be no need of Mullahs to travel from village to village."

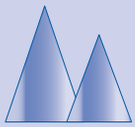
"When you first came to me you said you wanted to 'learn how to interpret' what you see as you travel through the world. Your confusion is simple. To interpret and to state Truths are two quite different things."

Having finished his story Halcolm smiled at the attentive youths. "Go, my children. Seek what you will, do what you must."

—From Halcolm's *Evaluation Parables*

Chapter Preview

This chapter concludes the book by addressing ways to enhance the quality and credibility of qualitative analysis. Module 76 discusses and demonstrates analytical processes for enhancing credibility by systematically engaging and questioning the data. Module 77 presents four triangulation processes for enhancing credibility. Modules 78 and 79 present alternative and competing criteria for judging the quality of qualitative studies. Module 80 discusses how and why the credibility of the inquirer is critical to the overall credibility of qualitative findings. Module 81 examines core issues of generalizability, extrapolations, transferability, generating principles, and harvesting lessons. Module 82 concludes the chapter and the book by addressing philosophy of science issues related to the credibility and utility of qualitative inquiry.



Analytical Processes for Enhancing Credibility: Systematically Engaging and Questioning the Data

The credibility of qualitative inquiry depends on four distinct but related inquiry elements:

1. *Systematic, in-depth fieldwork* that yields high-quality data
2. *Systematic and conscientious analysis of data* with attention to issues of credibility
3. *Credibility of the inquirer*, which depends on training, experience, track record, status, and presentation of self
4. *Readers' and users' philosophical belief in the value of qualitative inquiry*—that is, a fundamental appreciation of naturalistic inquiry, qualitative methods, inductive analysis, purposeful sampling, and holistic thinking (indeed, all 12 core qualitative strategies presented in Exhibit 2.1, pp. 46–47)

The first of the elements that determine credibility, *systematic, in-depth fieldwork that yields high-quality data*, was covered in Chapter 5 (purposeful qualitative designs), Chapter 6 (in-depth fieldwork and rich observational data), and Chapter 7 (high-quality, skillful interviewing).

This module and the next focus on the remaining three elements of quality: systematic and conscientious analysis of data. Module 80 discusses credibility of the inquirer, and Module 82 examines readers' and users' philosophical belief in the value of qualitative inquiry.

Strategies for Enhancing the Credibility of Analysis

> **Chance favors the prepared mind.**
>
> —Louis Pasteur (1822–1895)
> French microbiologist (known as the "father of microbiology") who discovered the process for pasteurizing milk, named after him

Chapter 8 presented analytical strategies for coding qualitative data, identifying patterns and themes, creating typologies, determining substantive significance, and reporting findings. However, at the heart of much controversy about qualitative findings are doubts about the nature of qualitative analysis because it is so judgment dependent. Statistical analysis follows formulas and rules, while, at the core, qualitative analysis depends on the insights, conceptual capabilities, and integrity of the analyst. Qualitative analysis is driven by the capacity for astute pattern recognition from beginning to end. Staying open to the data, for example, involves aggregating and integrating the data around a particular expected pattern while also watching for unexpected patterns. This process is epitomized in health research by the scientist working on one problem who suddenly notices a pattern related to a quite different problem—and thus discovers *Viagra*; as Pasteur explained when he was asked how he happened to discover how to stop bacterial contamination of milk, "Chance favors the prepared mind." Here, then, are some techniques that prepare the mind for insight while also enhancing the credibility of the resulting analysis.

Integrity in Analysis: Generating and Assessing Alternative Conclusions and Rival Explanations

One barrier to credible qualitative findings stems from the suspicion that the analyst has shaped findings according to his or her predispositions and biases. Being able to report that you engaged in a systematic and conscientious search for alternative themes, divergent patterns, and rival explanations enhances credibility, not to mention that it is simply good analytical practice and the very essence of being rigorous in analysis. This can be done both inductively and logically. Inductively, it involves looking for other ways of organizing the data that might lead to different findings. Logically, it means thinking about other logical possibilities and then seeing if those possibilities can be supported by the data. When considering rival organizing schemes and competing explanations, your mind-set should not be one of attempting to disprove the alternatives; rather, *you look for data that support alternative explanations*.

In evaluation of a training program for chronically unemployed men of color, we conducted case studies of a group of successes. The program model was based on training in both hard skills (e.g., machine tooling, keyboarding, welding, and accounting) and soft skills

(showing up to work on time, dressing appropriately, and respecting supervisors and coworkers). The cases studied validated the importance of both kinds of skills, but an additional explanation emerged in later cases, namely, that the program experience and peer support led to an identity shift: Successful trainees began to think of themselves as capable of holding a job. They were used to being labeled as "losers." The opportunity to think of themselves as "winners" involved more than acquiring "soft skills." It involved a shift in identity. We went back to earlier cases to find out if that phenomenon was evident there as well. It was, as was evidence for how that shift in identity occurred. Might this change be simply a function of participants being older by the time they entered this particular program (a maturation effect)? No, the change was evident in younger participants as well as older ones. We continued in this fashion, looking for alternative explanations and checking them out against the case data.

Failure to find strong supporting evidence for alternative ways of presenting data or contrary explanations helps increase confidence in the initial, principal explanation you generated. Comparing alternative patterns will not typically lead to clear-cut "yes there is support" versus "no there is no support" kinds of conclusions. You're searching for *the best fit*, the preponderance of evidence. This requires assessing the weight of evidence and looking for those patterns and conclusions that fit the preponderance of data. Keep track of and report alternative classification systems, themes, and explanations that you considered and "tested" during data analysis. This demonstrates intellectual integrity and lends considerable credibility to the final set of findings and explanations offered. Analysis of rival explanations in case studies is analogous to counterfactual analysis in experimental designs.

Searching for and Analyzing Negative or Disconfirming Evidence and *Case*s

Closely related to testing alternative constructs is the search for and analysis of *negative cases*. Where patterns and trends have been identified, our understanding of those patterns and trends is increased by considering the instances and cases that do not fit within the pattern. These may be exceptions that illuminate the boundaries of the pattern. They may also broaden understanding of the pattern, change the conceptualization of the pattern, or cast doubt on the pattern altogether.

> In qualitative analysis you need to keep analyzing the data to check any explanations and generalizations that you wish to make, to ensure that you have not missed anything that might lead you to question their applicability. Essentially this means looking for negative or deviant cases—situations and examples that just do not fit the general points you are trying to make. However, the discovery of negative cases or counter-evidence to a hunch in qualitative analysis does not mean its immediate rejection. You should investigate the negative cases and try to understand why they occurred and what circumstances produced them. As a result, you might extend the idea behind the code to include the circumstances of the negative case and thus extend the richness of your coding. (Gibbs, 2007, p. 96)

In the Southwest Field Training Project involving wilderness education, virtually all participants reported significant "personal growth" as a result of their participation in the wilderness experiences; however, the two people who reported "no change" provided particularly useful insights into how the program operated and affected participants. These two had crises going on back home that limited their capacity to "get into" the wilderness experiences. The project staff treated the wilderness experiences as fairly self-contained, closed-system experiences. The two negative cases opened up thinking about "baggage carried in from the outside world," "learning-oriented mind-sets," and a "readiness" factor that subsequently affected participant selection and preparation.

Negative cases also provide instructive opportunities for new learning in formative evaluations. For example, in a health education program for teenage mothers where the large majority of participants complete the program and show knowledge gains, an important component of the analysis should include examination of reactions from dropouts, even if the sample is small for the dropout group. While the small proportion of dropouts may not be large enough to make a difference in a statistical analysis, qualitatively the dropout feedback may provide critical information about a niche group or a specific subculture, and/or clues to program improvement.

No specific guidelines can tell you how and how long to search for negative cases or how to find alternative constructs and hypotheses in qualitative data. Your obligation is to make an "assiduous search . . . until no further negative cases are found" (Lincoln & Guba, 1986, p. 77). You then report the basis for the conclusions you reach about the significance of the negative or deviant cases.

Readers of a qualitative study will make their own decisions about the plausibility of alternate explanations and the reasons why deviant cases do not fit within dominant patterns. But I would note that the

SIDEBAR

ADVOCACY–ADVERSARY ANALYSIS

In 1587, the Roman Catholic Church created advocacy–adversary roles to test the validity of evidence in support of the canonization process for elevating someone to sainthood. The Devil's Advocate *(Latin:* advocatus diaboli*) in this process (officially designated the Promoter of the Faith) was a canon lawyer whose job was to argue against the canonization by presenting doubts about or holes in the evidence, for example, to argue that any miracles attributed to the candidate were unsubstantiated or even fraudulent. The Devil's Advocate opposed God's Advocate, whose job was to present evidence supporting and make the argument in favor of canonization. This advocacy–adversary process endured until 1983, when it was abolished by Pope John Paul II as overly adversarial and contentious.*

Advocacy–Adversary Analysis in Evaluation

A formal and forced approach to engaging rival conclusions draws on the legal system's reliance on opposing perspectives battling it out in the courtroom. The advocacy-adversary model suggested by Wolf (1975) developed in response to concerns that evaluators could be biased in their conclusions. Also called the *Judicial Model of Evaluation* (Datta, 2005), to balance possible evaluator biases, two teams engage in debate. The *advocacy team* gathers and presents information that supports the proposition that the program is effective; the *adversary team* gathers information that supports the conclusion that the program ought to be changed or terminated.

Some years ago, I served as the judge for what would constitute admissible evidence in an advocacy–adversary evaluation of an innovative education program in Hawaii. The task of the advocacy team was to gather and present data supporting the proposition that the program was effective and ought to be continued. The adversaries were charged with marshalling all possible evidence demonstrating that the program ought to be terminated. When I arrived on the scene, I immediately felt the exhilaration of the competition. I wrote in my journal,

> No longer staid academic scholars, these are athletes in a contest that will reveal who is best; these are lawyers prepared to use whatever means necessary to win their case. The teams have become openly secretive about their respective strategies. These are experienced evaluators engaged in a battle not only of data but also of wits.

As the two teams prepared their final reports, a concern emerged among some about the narrow focus of the evaluation. The summative question concerned whether the program should be continued or terminated. Education officials were asking how to improve the program without terminating it. Was it possible that a great amount of time, effort, and money was directed at answering the wrong question? Was it appropriate to force the data into a simple save-it-or-scrap-it choice? In fact, middle-ground positions were more sensible. But the advocacy–adversary analytical process design obliged opposing teams to do battle on the unembellished question of whether to maintain or terminate a program. A systematic assessment of strengths and weaknesses, with ideas for improvement, gave way to an all-good, all-bad framing, and that's how the results were presented (Patton, 2008, pp. 142–143).

The weakness of the advocacy–adversary approach is that it emphasizes contrasts and opposite conclusions, to the detriment of appreciating and communicating nuances in the data and accepting and acknowledging genuine and meaningful ambiguities. *Advocacy–adversary analysis* forces data sets into combat with each other. Such oversimplification of complex and multifaceted findings is a primary reason why advocacy–adversary evaluation is rarely used (in addition to being expensive and time-consuming). Still, it highlights the importance of engaging in some systematic analysis of alternative and rival conclusions, and as one approach (but not the only one) to testing conclusions, it can be useful and revealing.

Practical Analytical Variations on a Theme

1. A variation of the overall advocacy–adversary approach would be to arbitrarily create advocacy and adversary teams *only* during the analysis stage so that both teams work with the same set of data but each team organizes and interprets those data to support different and opposite conclusions, including identifying ambiguous findings.
2. Another variation would be for a lone analyst to organize data systematically into *pro* and *con* sets of evidence to see what each yielded.

section of the report that involves exploration of alternative explanations and consideration of why certain cases do not fall into the main pattern can be among the most interesting sections of a report to read. When well written, this section of a report reads something like a detective study in which the analyst (detective) looks for clues that lead in different directions and tries to sort out which direction makes the most sense given

SIDEBAR

ANALYTIC INDUCTION: HYPOTHESIS TESTING WITH NEGATIVE CASES

Analytic induction emphasizes giving special attention to negative or deviant cases for testing propositions that should, based on the theory being examined, apply to all cases that have been sampled in the design to manifest the phenomenon of interest. Analytic induction works through one case at a time. If the case data fit the hypothesis, the inductive analyst takes up the next case. If a case isn't consistent with the hypothesis—that is, it is a negative or deviant case—then the hypothesis is revised or the case is rejected as not actually relevant to the phenomenon being studied. The analytical focus is examining the extent to which *every case* confirms the hypothesis and to either refine the hypothesis or the statement of the problem to account for all cases. No cases can be ignored. All must be accounted for and used in the analysis.

Here's an example of testing a hypothesis about the effect of mother–daughter relationships on anorexia. The proposition being tested was "*If* mother was critical of daughter's body image *and* mother–daughter relationship was strained *and* daughter experiences weight loss, *then* count that as an example of mother' s negative influence on daughter's self-image." Once particular interviews were identified as containing the codes identified in the hypothesis, the qualitative data from interviews and cases could be examined to determine whether support for this causal interpretation could be justified for each case (Hesse-Biber & Dupuis, cited in Silverman & Marvasti, 2008, p. 252). The rigor of this approach is that finding even a single disconfirming case disconfirms the hypothesis requiring either refinement or reformulation, for the goal is to identify and confirm a generalizable, universal, causal explanation for the phenomenon of interest (Flick, 2007a, p. 30; Schwandt, 2007, p. 6).

the clues (data) that are available. Such writing adds credibility by showing the analyst's authentic search for what makes most sense rather than marshalling all the data toward a single conclusion. Indeed, the whole tone of a report feels different when the qualitative analyst is willing to openly consider other possibilities than those finally settled on as most reasonable in accordance with the preponderance of evidence. Compare the approach of weighing alternatives with the report where all the data lead in a single-minded fashion, in a rising crescendo, toward an overwhelming presentation of a single point of view. Perfect patterns and omniscient explanations are likely to be greeted skeptically—and for good reason: The human world is not perfectly ordered, and human researchers are not omniscient. Humility can do more than certainty to enhance credibility. Dealing openly with the complexities and dilemmas posed by negative cases is both intellectually honest and politically strategic.

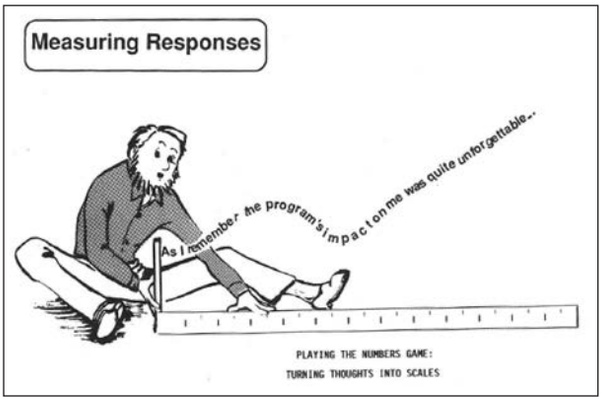
Avoid the Numbers Game

Philosopher of science Thomas H. Kuhn (1970), having studied extensively the value systems of scientists, observed that "the most deeply held values concern predictions" and "quantitative predictions are preferable to qualitative ones" (pp. 184–185). The methodological status hierarchy in science ranks "hard data" above "soft data," where "hardness" refers to the precision of statistics. Qualitative data can carry the stigma of "being soft."This carries over into the public arena, especially in the media and among policymakers, creating what has been called the tyranny of numbers (Eberstadt, 1995).

How can one deal with a lingering bias against qualitative methods? A starting point is helping people understand that qualitative methods are not weaker or softer than quantitative approaches. Qualitative methods are *different*. Making the case for the value of qualitative inquiries involves being able to communicate the particular strengths of qualitative methods (Chapters 1 and 2) and the kinds of evaluation and other applications for which qualitative data are especially appropriate (Chapter 4). But those understandings can only open the door to dialogue. The fact is that numbers have a special allure in modern society. Statistics are seductive—so precise, so clear. Numbers convey that sense of precision and accuracy, even if the measurements that yielded the numbers are relatively unreliable, invalid, and meaningless (e.g., see Hausman, 2000; Silver, 2012).

Quantitizing

Quantitizing, commonly understood to refer to the numerical translation, transformation, or conversion of qualitative data, has become a staple of mixed-methods research (Sandelowski, Voils, &



©2002 Michael Quinn Patton and Michael Cochran

Knafl, 2009, p. 208). Quantitized qualitative data are analyzed statistically, including using statistical significance tests (Collingridge, 2013).

There are different techniques by which quantitization may be achieved. Two common strategies are (1) dichotomizing and (2) counting. Dichotomizing refers to assigning a binary value (e.g., 0 and 1) to variables with two mutually exclusive and exhaustive categories, such as assigning "0" to participants who did not express a particular theme and "1" to participants who did express the theme. In contrast, counting involves calculating the number of themes expressed by each participant, as in the case of determining that a participant expressed two out of four themes in a study. Counting also includes calculating the number of qualitative codes assigned to specific themes, as in the case of determining that a participant expressed 10 qualitative codes associated with a theme (Collingridge, 2013, p. 82).

In Chapter 8, I devoted my MQP Rumination to why I consider this kind of quantizing to be generally a bad idea and advocated keeping qualitative analysis qualitative (see pp. 557–559). I won't repeat that argument here. Still, it strikes me as a worrisome trend. What's driving it? Partly, it's simply the cultural and political allure of numbers. But there's more.

> Pragmatic and ecumenical impulses, and the advent of computerized software programs to manage both qualitative and quantitative data, have served to promote a largely technical view of quantitizing. Moreover, the rhetorical appeal of numbers—their cultural association with scientific precision and rigor—has served to reinforce the necessity of converting qualitative into quantitative data.

A systematic literature review of quantitizing studies—that is, studies featuring quantitative analysis of qualitative interviews—shows the widespread nature of the phenomenon and some of the problems that arise, especially applying statistics to small sample sizes. Quantitative analyses of qualitative data are done to disaggregate results by background characteristics of participants (cross-tabs and correlations), to statistically test hypotheses, and to determine the prevalence of themes. But the overall problem is precisely what one would expect: "The conversion of the qualitative information to frequency counts has reduced the rich interpretation of people's experience that was expressed through their interviews" (Fakis, Hilliam, Stoneley, & Townend, 2014, p. 156). That is the crux of the issue, as is replacing a determination of substantive significance with the safe fallback position of replying on statistical significance. Moreover, those engaged in quantiziting seem oblivious to the issues involved.

> Typically glossed, however, are the foundational assumptions, judgments, and compromises involved in converting qualitative into quantitative data and whether such conversions advance inquiry. . . . Such conversions "are by no means transparent, uncontentious, or apolitical" (Love, Pritchard, Maguire, McCarthy, & Paddock, 2005, p. 287; Sandelowski, Voils, & Knafl, 2009, p. 28).

Substantive Significance Trumps Statistical Significance

The point, however, is not to be anti-numbers. The point is to be *pro-meaningfulness*.

I'm not numbers phobic. I have used numbers regularly in titling exhibits throughout this book:

> **Exhibit 8.1** Twelve Tips for Ensuring a Strong Foundation for Qualitative Analysis (pp. 522–523).
>
> **Exhibit 8.10** Ten Types of Qualitative Analysis (see pp. 551–552).
>
> **Exhibit 9.1** Ten Systematic Analysis Strategies to Enhance Credibility and Utility (pp. 659–660).
>
> Module 77 presents four triangulation processes for enhancing credibility.

When there is something meaningful to be counted, then count. As sample sizes increase, especially in mixed-methods studies, quantizing is likely to become even more pervasive. One study in the systematic review of quantizing articles had a sample size of 400 (Fakis et al., 2014, p. 146). Such studies will quantitize and do so appropriately. Weaver-Hightower (2014) studied political influence by reviewing public policy documents; from 1,459 transcript pages, he coded 2,294 unique arguments and relied heavily on quantitative analysis. That's understandable and appropriate, though reporting the results to two decimal places, "the average agreement score was 5.22%" (p. 125), illustrates the allure of pretentious precision. Or maybe just habit.

So while I advocate keeping qualitative analysis qualitative and focusing on substantive significance when interpreting findings, this is no hard-and-fast rule (my Chapter 8 MQP Rumination notwithstanding). Do what is appropriate. It doesn't make sense to report percentages in a sample of 10 interviewees; it does make sense with a sample of 400. By knowing the strengths and weaknesses of both quantitative and qualitative data, you can help those with whom

SIDEBAR

CONSTANT COMPARISON

A lot of qualitative analysis involves comparisons: comparing cases, comparing quotations, comparing observations, and comparing findings in others studies with your own findings.

The point about these comparisons is that they are constant; they continue throughout the period of analysis and are used not just to develop theory and explanations but also to increase the richness of description in your analysis and thus ensure that it closely captures what people have told you and what happened.

There are two aspects to this constant process:

1. Use the comparisons to check the *consistency and accuracy* of application of your codes, especially as you first develop them. Try to ensure that the passages coded the same way are actually similar. But at the same time, keep your eyes open for ways in which they are different. Filling out the detail of what is coded in this way may lead you to further codes and to ideas about what is associated with any variation. This can be seen as a circular or iterative process. Thus, develop your code, check for other occurrences in your data, compare these with the original, and then revise your coding (and associated memos) if necessary.
2. Look explicitly for *differences and variations* in the activities, experiences, actions and so on that have been coded. In particular, look for variation across cases, settings and events (Gibbs, 2007, p. 96).

Constant comparison is an ongoing analysis of similarities and differences: What things go together in the data? What things are different? What explains these similarities and differences? What are the implications for your overall inquiry purpose and conclusions?

Design Checks: Keeping Methods and Data in Context

One issue that can arise during analysis is concerns about how design decisions affect results. For example, purposeful sampling strategies provide a limited number of cases for examination. When interpreting findings, it becomes important to reconsider how design constraints may have affected the data available for analysis. This means considering the rival methodological hypothesis that the findings are due to methodological idiosyncrasies.

By their nature, qualitative findings are highly context and case dependent. Three kinds of sampling limitations typically arise in qualitative research designs:

1. There are limitations in the situations (critical events or cases) that are sampled for observation (because it is rarely possible to observe all situations even within a single setting).

you dialogue focus on really important questions rather than, as sometimes happens, focusing primarily on how to generate numbers. The really important questions are about what the findings mean. A single illuminative case or interview may be more substantively meaningful and insightful than 20 routine cases. That 5% level of insight is not a reason to pay more attention to the 95% degree of mediocrity just because there's more of it. Information-rich cases stand out not because there are lots of them but precisely because they are so rare—and rich with revelation (the very definition of being information-rich). Rare, precious gems are valued over widely available (and less expensive), semiprecious stones for the same reason. Qualitative analysis must include the analytical insight to distinguish signal from noise and valuable insights from commonplace ones.

Summary of Strategies for Systematically Analyzing Qualitative Data to Enhance Credibility

Qualitative analysis aims to make sense of qualitative data: detecting patterns, identifying themes, answering the primary questions framing the study, and presenting substantively significant findings. In this chapter, we've been looking at ways of enhancing the credibility of findings by deepening the analysis, reexamining initial findings, and continuously working back and forth between the findings and the data to validate findings against data. Exhibit 9.1 summarizes the analytical techniques we've just covered and looks ahead to the four kinds of triangulation I'll present and discuss in the next module (Items 7–10 in Exhibit 9.1).

2. There are limitations from the time periods during which observations took place—that is, constraints of temporal sampling.
3. The findings will be limited based on selectivity in the people who were sampled for observations or interviews, or selectivity in document sampling.

In reporting how purposeful sampling decisions affect findings, the analyst returns to the reasons for having made the initial design decisions. Purposeful sampling involves studying information-rich cases in depth and detail to understand and illuminate important cases rather than generalizing from a sample to a population (see Chapter 5). For instance, sampling and studying highly successful and unsuccessful cases in an intervention yields quite different results from studying a "typical" case or a mix of cases. People unfamiliar with purposeful samples may think of small, purposeful samples as "biased," a perception that undermines credibility in their minds. In communicating findings, then, it becomes important to emphasize that the issue is not one of dealing with a distorted or biased sample but rather one of clearly delineating the purpose, strengths, and limitations of the sample studied—and therefore being careful about not inappropriately extrapolating the findings to other situations, other time periods, and other people—a caution we'll return to later in this chapter. Reporting both methods and results in their proper contexts will avoid many controversies that result from yielding to the temptation to overgeneralize from purposeful samples. *Keeping findings in context is a cardinal principle of qualitative analysis. Design decisions are context for analysis.*

The wise fool in Sufi tales, Mulla Nasrudin, was once called on to make this point to his monarch. Although he was supposed to be a wise man, Nasrudin was accused of being illiterate. Nagged to action by skeptics, the monarch decided to test him.

"Write something for me, Nasrudin," said the king.

"I would willingly do so, but I have taken an oath never to write so much as a single letter again," replied Nasrudin.

"Well, write something in the way in which you used to write before you decided not to write, so that I can see what it was like."

"I cannot do that, because every time you write something, your writing changes slightly through practice. If I wrote now, it would be something written for now."

"Then," addressing the crowd, the king commanded: "Bring me an example of Nasrudin's writing, anyone who has something he's written."

Someone brought a terrible scrawl that Nasrudin had once written to him.

"Is this your writing?" asked the monarch.

"No," said Nasrudin. "Not only does writing change with time, but reasons for writing change. You are now showing a piece of writing done by me to demonstrate to someone how he should *not* write." (Shah, 1973, p. 92)

EXHIBIT 9.1 Ten Systematic Analysis Strategies to Enhance Credibility and Utility

1. *Generate and assess alternative conclusions and rival explanations.* Don't settle quickly on initial conclusions. Go back to the data. What are other ways of explaining what you've found? Look for the explanation that best fits the preponderance of evidence.

2. *Advocacy–adversary analysis uses a debate format for testing the viability of conclusions.* What are the evidence and arguments that support your conclusions? What are the contrary evidence and counterarguments? Get another analyst to play the "Devil's Advocate" role, or switch back and forth in advocacy and adversary roles yourself. The aim is to surface doubts and weaknesses as well as build on strengths and confirm solid conclusions.

3. *Search for and analyze negative or disconfirming evidence and cases.* There are "exceptions that prove the rule" and exceptions that question the rule. In either case, look for and learn from exceptions to the patterns you've identified.

4. *Make constant comparison your constant companion.* All analysis is ultimately comparative. You compare the data that fit into a category, pattern, or theme with the data that don't fit. You compare alternative explanations, conclusions, and chains of evidence. Compare and contrast. Then compare and contrast some more.

5. *Keep analysis connected to purpose and design.* When deeply enmeshed in cataloguing, classifying, and comparing the trees in your qualitative data—that is, the depth and details of rich, thick qualitative data—change perspectives now and again to see the forest—that is, reconnect with the big picture. Purpose drives design. Purpose and design drive data collection. Purpose, design, and the data collected, in combination, drive analysis. Make sure that your analysis is serving the purpose of the inquiry. A well-chosen, thoughtful design will have anticipated how analysis would unfold. Keep those linkages in mind so that analysis doesn't become isolated from the inquiry's overall purpose and context.

   Keeping findings in context is a cardinal principle of qualitative analysis.

6. *Keep qualitative analysis qualitative.* Paraphrasing poet Dylan Thomas, do not go gently into that numerical night. Quantitize thoughtfully, carefully, and even reluctantly. Do so when it's appropriate and enhances understanding, all the while aware of the allure of numbers and the danger of losing the richness of qualitative data in the parsimony of numerical reduction.

7. *Integrate and triangulate diverse sources of qualitative data: interviews, observations, document analysis.* Any single source of data has strengths and weaknesses. Consistency of findings across types of data increases confidence in the confirmed patterns and themes. Inconsistency across types of data invites questions and reflection about why certain methods produced certain findings.

8. *Integrate and triangulate quantitative and qualitative data in mixed-methods studies.* The logic of triangulation (see Item 7) applies in mixed-methods designs when the strengths and weaknesses of qualitative and quantitative data are used together to illuminate the inquiry.

9. *Triangulate analysts.* Having more than one pair of eyes look at and think about the data, identify patterns and themes, and test conclusions and explanations reduces concerns about the potential biases and selective perception of a single analyst.

10. *Undertake theory triangulation.* Look at the findings and conclusions through the lens of alternative theoretical frameworks. How would a symbolic interactionist interpret the data compared with a phenomenologist or realist? How would a behavioral psychologist interpret the findings compared with a humanistic psychologist? What does a mechanistic display reveal compared with a systems graphic? The point is not to conduct an endless set of such theoretical comparisons but to select only those theoretical frameworks most germane to your inquiry to see what the alternative perspectives yield by way of insight and explanation.

Four Triangulation Processes for Enhancing Credibility

> **By combining multiple observers, theories, methods and data sources, [researchers] can hope to overcome the intrinsic bias that comes from single-methods, single-observer, and single-theory studies.**
>
> —Norman K. Denzin (1989c, p. 307)

Chapter 5 on design discussed the benefits of using multiple data-collection techniques, a form of triangulation, to study the same setting, issue, or program. You may recall from that discussion that the term *triangulation* is taken from land surveying. Knowing a single landmark only locates you somewhere along a line in a direction from the landmark, whereas with two landmarks you can take bearings in two directions and locate yourself at their intersection. The notion of triangulating also works metaphorically to call to mind the world's strongest geometric shape—the triangle, which in its double alchemical form serves as the symbol for this chapter. The logic of triangulation is based on the premise that no single method ever adequately solves the problem of rival explanations. Because each method reveals different aspects of empirical reality and social perception, multiple methods of data collection and analysis provide more grist for the analytical mill. Combinations of interviewing, observation, and document analysis are expected in most fieldwork. Mixed qualitative–quantitative studies are increasingly valued as more credible than single-method studies. Studies that use only one method are more vulnerable to errors linked to that particular method (e.g., loaded interview questions, biased or untrue responses) than studies that use multiple methods, in which different types of data provide cross-data consistency checks.

Four Kinds of Analytical Triangulation

It is in data analysis that the strategy of triangulation really pays off, not only in providing diverse ways of looking at the same phenomenon, but in adding to credibility by strengthening confidence in whatever conclusions are drawn. Four kinds of triangulation can contribute to the verification and validation of qualitative analysis:

1. ***Triangulation of qualitative sources:*** Checking out the consistency of different data sources within the same method (consistency across interviewees)
2. ***Mixed qualitative–quantitative methods triangulation:*** Checking out the consistency of findings generated by different data collection methods
3. ***Analyst triangulation*:** Using multiple analysts to review findings
4. ***Theory/perspective triangulation*:** Using multiple perspectives or theories to interpret data

By triangulating with multiple data sources, methods analysts, and/or theories, qualitative analysts can make substantial strides in overcoming the skepticism that greets singular methods, lone analysts, and single-perspective interpretations.

Interpreting Triangulation Results: Making Sense of Conflicting and Inconsistent Patterns

A common misconception about triangulation involves thinking that the purpose is to demonstrate that different data sources or inquiry approaches yield essentially the same result. The point is to *test for* such consistency. Different kinds of data may yield somewhat different results because different types of inquiry are sensitive to different real-world nuances. Thus, *understanding inconsistencies in findings across different kinds of data can be illuminative and important.* Finding such inconsistencies ought not to be viewed as weakening the credibility of results but, rather, as offering opportunities for deeper insight into the relationship between inquiry approach and the phenomenon under study. I'll comment briefly on each of the four types of triangulation.

1. Triangulation of Qualitative Data Sources

> **Four kinds of persons: zeal without knowledge; knowledge without zeal; neither knowledge nor zeal; both zeal and knowledge.**
>
> —Pascal, *Pensées*

Four kinds of qualitative triangulation: interviews with observations; interviews with documents; observations with documents; and interviews from multiple sources with observations of diverse events and documents of many kinds.

—Halcolm, *Qualitative Pensées*

Triangulation of data sources within and across different qualitative methods means comparing and cross-checking the consistency of information derived at different times and by different means from interviews, observations, and documents. It can include

- comparing observations with interviews;
- comparing what people say in public with what they say in private;
- checking for the consistency of what people say about the same thing over time;
- comparing the perspectives of people from different points of view—for example, in an evaluation, triangulating staff views, participants' views, funder views, and views expressed by people outside the program; and
- checking interviews against program documents and other written evidence that can corroborate what interview respondents report.

Quite different kinds of data can be brought together in a case study to illuminate various aspects of a phenomenon. In a classic evaluation of an innovative educational project, historical program documents, in-depth interviews, and ethnographic participant observations were triangulated to illuminate the roles of powerful actors in supporting adoption of the innovation (Smith & Kleine, 1986). The evaluation of the Paris Declaration on development aid triangulated interviews with a variety of key informants, government reports, donor agency reports, and observations of donor–recipient decision-making meetings (Wood et al., 2011).

Maxwell (2012) is especially insightful about the interrelationship of interview and observation data in qualitative inquiry and analysis.

> One belief that inhibits triangulation is the widespread (though often implicit) assumption that observation is mainly useful for describing behavior and events, while interviewing is mainly useful for obtaining the perspectives of actors. It is true that the immediate result of observation is description, but this is equally true of interviewing: The latter gives you a description of what the informant said, not a direct understanding of their perspective. Generating an interpretation of someone's perspective is inherently a matter of inference from descriptions of their behavior (including verbal behavior), whether the data are derived from observations, interviews, or some other source such as written documents.
>
> While interviewing is often an efficient and valid way of understanding someone's perspective, observation can enable you to draw inferences about this perspective that you couldn't obtain by relying exclusively on interview data. . . . For example, watching how a teacher responds to boys' and girls' questions in a science class may provide a much better understanding of the teacher's actual views about gender and science than what the teacher says in an interview.
>
> Conversely, although observation often provides a direct and powerful way of learning about people's behavior and the context in which this occurs, interviewing can also be a valuable way of gaining a description of actions and events—often the only way, for events that took place in the past or to which you can't gain observational access. Interviews can provide additional information that was missed in observation, and can be used to check the accuracy of the observations. However, in order for interviews to be useful for this purpose, you need to ask about *specific* events and actions rather than posing questions that elicit only generalizations or abstract opinions. . . . In both of these situations, triangulation of observations and interviews can provide a more complete and accurate account than either could alone. (pp. 106–107)

Triangulation of data sources within qualitative methods may not lead to a single, totally consistent picture. The point is to study and understand when and why differences appear. The fact that observational data produce different results from interview data does not mean that either or both kinds of data are "invalid," although that may be the case. More likely, it means that different kinds of data have captured different things and so the analyst attempts to understand the reasons for the differences. Either consistency in overall patterns of data from different sources or reasonable explanations for differences in data from divergent sources can contribute significantly to the overall credibility of findings.

SIDEBAR

ETHNOGRAPHIC TRIANGULATION

In ethnographic research practice, triangulation of data sorts and methods and of theoretical perspectives leads to extended knowledge potentials, which are fed by the convergences, and even more by the divergences, they produce.

> As in other areas of qualitative research, triangulation in ethnography is a way of promoting quality of research. . . . Good ethnographies are characterized by flexible and hybrid use of different ways of collecting data and by a prolonged engagement in the field. As in other areas of qualitative research, triangulation can help reveal different perspectives on one issue in research such as knowledge about and practices with a specific issue. Thus, triangulation is again a way to promote quality of qualitative research in ethnography also and more generally a productive approach to managing quality in qualitative research. (Flick, 2007b, p. 89)

2. Mixed-Methods Triangulation: Integrating Qualitative and Quantitative Data

> 'Tis not the many oaths that makes the truth,
>
> But the plain single vow that is vow'd true.
>
> —William Shakespeare (written 1604–1605)
> Diana in *All's Wells That Ends Well*

Mixed-methods triangulation often involves comparing and integrating data collected through some kind of qualitative methods with data collected through some kind of quantitative method. Such efforts flow from a pragmatic approach to mixed-methods analysis that assumes potential compatibility and seeks to discover the degree and nature of such compatibility (Tashakkori & Teddlie, 1998; Teddlie & Tashakkori, 2003, 2011). This is seldom straightforward because certain kinds of questions lend themselves to qualitative methods (e.g., developing hypotheses or theory in the early stages of an inquiry, understanding particular cases in depth and detail, getting at meanings in context, and capturing changes in a dynamic environment), while other kinds of analyses lend themselves to quantitative approaches (e.g., generalizing from a sample to a population, testing hypotheses for statistical significance, and making systematic comparisons on standardized criteria). Thus, it is common that quantitative methods and qualitative methods are used in a complementary fashion to answer different questions that do not easily come together to provide a single, well-integrated picture of the situation.

Given the varying strengths and weaknesses of qualitative versus quantitative approaches, the researcher using different methods to investigate the same phenomenon should not expect that the findings generated by those different methods will automatically come together to produce some nicely integrated whole. Indeed, the evidence is that one ought to expect initial conflicts in findings from qualitative and quantitative data and expect those findings to be received with varying degrees of credibility. It is important, then, to consider carefully what each kind of analysis yields and thereby giving different interpretations the chance to arise, with each considered on its merits, before favoring one result over the other based on methodological biases.

Critical Multiplism as an Analytical Strategy

Critical multiplism is a research strategy that advocates designing packages of imperfect methods and theories in a manner that minimizes the respective and inevitable biases of each. Multiplism, applied to analysis, acknowledges that any analysis can usually be conducted in any one of several ways, but in many cases, no single way is known to be uniformly the best. Under such circumstances, a multiplist advocates making heterogeneous those aspects of analysis about which uncertainty exists, so that the task is conducted in several different ways, each of which is subject to different biases.

> *Critical* refers to rational, empirical, and social efforts to identify the assumptions and biases present in the options chosen. Putting the two concepts together, we can say that the central tenet of critical multiplism is this: When it is not clear which of several defensible options for a scientific task is least biased, we should select more than one, so that our options reflect different biases, avoid constant biases, and leave no plausible bias overlooked. (Shadish, 1993, p. 18)

When multiple analytical approaches yield similar results across different analytical biases, confidence in the resulting findings is increased. If different results occur when the analysis is done in different ways, then we have to try to explain the differences.

Different Findings From Different Methods

In a classic article, Shapiro (1973) described in detail her struggle to resolve basic differences between qualitative data and quantitative data in her study of Follow Through Classrooms; she eventually concluded that some of the conflicts between the two kinds of data were the result of measuring different things, although the ways in which different things were measured were not immediately apparent until she worked to sort out the conflicting findings. She began with greater trust in the data derived from quantitative methods and ended by believing that the most useful information came from the qualitative data.

Another pioneering article, by M. G. Trend (1978) of ABT Associates, has become required reading for anyone becoming involved in a team project that will involve collecting and analyzing both qualitative and quantitative data, where different members of the team have responsibilities for different kinds of data. The Trend study involved an analysis of three social experiments designed to test the concept of using direct-cash housing allowance payments to help low-income families obtain decent housing on the open market. The analysis of qualitative data from a participant observation study produced results that were at variance with those generated by analysis of quantitative data. The credibility of the qualitative data became a central issue in the analysis.

> The difficulty lay in conflicting explanations or accounts, each based largely upon a different kind of data. The problems we faced involved not only the nature of observational versus statistical inferences, but two sets of preferences and biases within the entire research team. . . .
>
> Though qualitative/quantitative tension is not the only problem which may arise in research, I suggest that it is a likely one. Few researchers are equally comfortable with both types of data, and the procedures for using the two together are not well developed. The tendency is to relegate one type of analysis or the other to a secondary role, according to the nature of the research and the predilections of the investigators. . . . Commonly, however, observational data are used for "generating hypotheses," or "describing process." Quantitative data are used to "analyze outcomes," or "verify hypotheses." I feel that this division of labor is rigid and limiting. (Trend, 1978, p. 352)

Early Efforts at Quantitative–Qualitative Triangulation

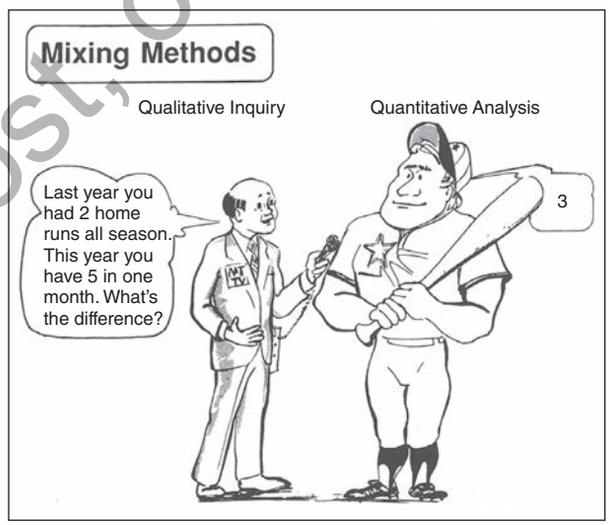
Anthropologists participating in teams in which both quantitative and qualitative data were being collected applied their inquiry skills to examine the nature of the experience in the 1970s. The problems they have shared were stark evidence that qualitative methods at that time were typically perceived as exploratory and secondary when used in conjunction with quantitative/experimental approaches. When qualitative data supported quantitative findings, that was the icing on the cake. When qualitative data conflicted with quantitative data, the qualitative data have often been dismissed or ignored (Society of Applied Anthropology, 1980).

A strategy of methods triangulation, then, doesn't magically put everyone on the same page. While valuing and endorsing triangulation, Trend (1978) suggested that

SIDEBAR

STRATEGY FOR ACHIEVING QUALITY IN MIXED-METHODS STUDIES

The quantitative researchers work side by side every step of the way as full members of the case study team, bringing the analytic rigor of their quantitative frameworks to bear on case study and observation design, data collection, analysis, integration with other methods, and reporting. The qualitative researchers, in turn, are full members of the quantitative team (analysis of administrative data, survey research, and time series assessments), bringing their own rigor to survey designs, data reduction decisions, and interpretations. As a result, assumptions are more rigorously examined, methodological lacunae more clearly (and early) identified, and the team leaders become sufficiently methodologically multilingual so that they can discuss both qualitatively and quantitatively based findings with equal confidence (Datta, 2006, p. 427).



> we give different viewpoints the chance to arise, and postpone the immediate rejection of information or hypotheses that seem out of joint with the majority viewpoint. Observationally derived explanations are particularly vulnerable to dismissal without a fair trial. (pp. 352–353)

From Separation to Integration

Qualitative and quantitative data can be fruitfully combined to elucidate complementary aspects of the same phenomenon. For example, a community health indicator (e.g., teenage pregnancy rate) can provide a general and generalizable picture of an issue, while case studies of a few pregnant teenagers can put faces on the numbers and illuminate the stories behind the quantitative data; this becomes even more powerful when the indicator is broken into categories (e.g., those under the age of 15, those 16 and above), with case studies illustrating the implications of and rationale for such categorization.

SIDEBAR

A STORY OF MIXED-METHODS TRIANGULATION: TESTING CONCLUSIONS WITH MORE FIELDWORK

Economists Lawrence Katz and Jeffrey Liebman of Harvard, and Jeffrey R. Kling of Princeton, were trying to interpret data from a federal housing experiment that involved randomly assigning people to a program that would help them get out of the slums. The evaluation focused on the usual outcomes of improved school and job performance. However, to get beyond the purely statistical data, they decided to conduct interviews with residents in an inner-city poverty community.

Professor Lieberman commented to a *New York Times* reporter,

> I thought they were going to say they wanted access to better jobs and schools, and what we came to understand was their consuming fear of random crime; the need the mothers felt to spend every minute of their day making sure their children were safe. (Uchitelle, 2001, p. 4)

By adding qualitative, field-based interview data to their study, Kling, Liebman, and Katz (2001) came to a new and different understanding of the program's impacts and participants' motivations based on interviewing the people directly affected, listening to their perspectives, and including those perspectives in their analysis.

In essence, triangulation of qualitative and quantitative data constitutes a form of comparative analysis. The question is "What does each analysis contribute to our understanding?" Areas of convergence increase confidence in findings. Areas of divergence open windows to better understanding of the multifaceted, complex nature of a phenomenon. Deciding whether results have converged remains a delicate exercise subject to both disciplined and creative interpretation. Focusing on *the degree of convergence* rather than forcing a dichotomous choice—that the different kinds of data do or do not converge—yields a more balanced overall result.

Mixed-Methods Analysis and Triangulation in the Twenty-First Century

While difficulties still arise in triangulating and integrating qualitative and quantitative data, advances in mixed methods have propelled integrated analyses into the spotlight, especially in applied and interdisciplinary areas like policy analysis, program evaluation, environmental studies, international development, and global health. Where disciplinary barriers have yielded to genuine interdisciplinary engagement, traditional methodological divisions have yielded to collaboration and integration. Exhibit 8.27 (pp. 618–619) presented mixed-methods challenges and solutions. Exhibit 9.2 presents 10 developments that are making mixed-methods triangulation both valued and, increasingly, expected in applied social science.

3. Triangulation With Multiple Analysts

A third kind of triangulation is investigator or analyst triangulation—that is, using multiple as opposed to singular observers or analysts. This is the core of qualitative team research (Guest & MacQueen, 2008). Triangulating observers or using several interviewers helps reduce the potential bias that comes from a single person doing all the data collection and provides means of more directly assessing the consistency of the data obtained. Triangulating observers provides a check on potential bias in data collection.

A related strategy is *triangulating analysts*—that is, having two or more persons independently analyze the same qualitative data and compare their findings. In the traditional social science approach to qualitative inquiry, engaging multiple analysts and computing the interrater reliability among these different analysts is valued, even expected, as a means of establishing credibility of findings (Silverman & Marvasti, 2008, pp. 238–239).

EXHIBIT 9.2 Ten Developments Enhancing Mixed-Methods Triangulation

1. *Designs that are truly mixed-methods inquiries are demonstrating the value of systematic, planned triangulation.* Increased understanding of the strengths and weaknesses of qualitative and quantitative data has led to both the commitment and capacity to build on the strengths of each *at the design stage.*

2. *Asking integrating questions of the data supports triangulation.* Triangulation is most powerful when mixed-methods studies are designed for integration, which begins by asking the same questions of both methods and gathering both qualitative and quantitative data on those questions. That is happening at a level unprecedented in applied social science research and evaluation.

3. *Mixed-methods sampling strategies anticipate and facilitate triangulation.* Sampling with triangulation in mind is a collaborative strategy that anticipates and lays the foundation for mixed-methods analysis.

4. *Specific methods are incorporating mixed data intentionally to support triangulation.* Surveys ask both closed- and open-ended questions. Case studies collect both quantitative and qualitative data. Strong experimental designs gather both standardized intervention and quantitative effects data plus qualitative process data.

5. *Mixed methods are proving especially appropriate for studying complex issues.* Mixed-methods researchers are extending our understandings of how to understand complex social phenomena as well as how to use research to develop effective interventions to address complex social problems (Mertens, 2013; Patton, 2011).

6. *Team approaches are being created and implemented with mixed-methods skills and capabilities in mind.* High-quality mixed-methods designs often require teams because individuals lack the full skill set needed. Knowing how to form and manage such teams has advanced significantly as experience has accumulated about what to do—and what not to do (Guest & MacQueen, 2008; Morgan, 2014).

7. *Software supports mixed-methods data analysis and triangulation.* As data analysis software has become more sophisticated, flexible, and responsive to analysts' needs, techniques and processes for triangulation are becoming more common and easier to use.

8. *Resources available for mixed-methods designs and analysis have burgeoned.* The *Journal of Mixed Methods* began publishing in 2007, with an opening editorial by Abbas Taskhakkori and John Creswell proclaiming, "The New Era of Mixed Methods." This means that there are more outlets for publishing mixed-methods studies. *The Handbook of Mixed Methods* was published in 2003 (Tashakkori & Teddlie). Excellent mixed-methods texts provide guidance on the full process from designing mixed-methods studies to analyzing and triangulating mixed data (Bamberger, 2013; Bergman, 2008; Greene, 2007; Mertens, 1998; Mertens & Hesse-Biber, 2013; Morgan, 2014).

9. *Researchers are developing mixed skills, capabilities, and capacities—and being recognized and valued for their mixed-methods expertise.* In 2014, the International Association of Mixed Methods Research was launched and hailed as "a momentous development in mixed-methods research" (Mertens, 2014).

10. *Mixed-methods exemplars show what is possible.* Early experiences with qualitative–quantitative triangulation were mixed at best—and many were quite negative, as indicated in the cautionary tales reported preceding the exhibit. When I was doing earlier editions of this book, there were more bad examples and negative experiences than good and positive exemplars. That balance has shifted for all the reasons listed here. The momentum is building as funders of research and evaluation are coming to demand mixed-methods studies.

Here, however, is a perfect example of how different criteria for judging quality lead to different practices. In a lead editorial for the journal *Qualitative Health Research*, Janet Morse (1997) took on "the myth of inter-rater reliability" from a social constructionist perspective. She begins by distinguishing standardized interview formats from more flexible and open interview guide approaches. She acknowledges that interrater reliability may be acceptable when everyone is asked the same question in the same way (the preferred interviewing approach to meet traditional social science concerns about validity and reliability), but in the more adaptive, personalized, individualized, and flexible approach of interview guides and conversational interviewing, what constitutes coherent passages

for coding is more problematic and depends on the analyst's interpretive framework. Multiple analysts might still discuss what they see in the data, share insights, and consider what emerges from their different perspectives, but that's quite different from computing a statistical interrater reliability coefficient. (See the sidebar on "the myth of interrater reliability" for her full argument.)

SIDEBAR

PERFECTLY HEALTHY BUT DEAD: THE MYTH OF INTERRATER RELIABILITY

—Janet M. Morse (1997)

Qualitative researchers seem to have inherited a host of habits from quantitative researchers and have adopted them into the qualitative paradigm without considering the appropriateness of their purpose, rationale, or underlying assumptions. On the surface, these practices seem right, so they are unquestioningly maintained. One of these adopted habits is the practice of obtaining interrater reliability of coding decisions used in qualitative research when coding unstructured, interactive interviews.

The argument goes something like this: To be reliable, coding should be replicable. Replication is checked by duplication; if coding decisions are explicit and communicated to another researcher, that researcher should be able to make the same coding decisions as the first researcher. The result is reliable research. Right?

Wrong. Interrater reliability is appropriate with semistructured interviews, wherein all participants are asked the same questions, in the same order, and data are coded all at once at the end of the data collection period. But this does not hold for unstructured interactive interviews. Recall that unstructured, interactive interviews are used in research because the researcher does not know enough about the topic or its parameters to construct interview questions. With unstructured, interactive interviews, the researcher first assumes a listening stance and learns about the topic as she or he goes along. Thus, once the researcher has learned something about the phenomenon from the first few participants, the substance of the interview then changes and becomes targeted on another aspect of the phenomenon. Importantly, unlike semistructured interviews, all participants are not asked the same questions. Participants are used to verify the information learned in the first interviews and are encouraged both to speak from their own experience and to speak for others. Each interview may overlap with the others but may also have a slightly different focus and different content.

This notion, learning from participants as the study progresses, is crucial to the understanding of the fluid nature of coding unstructured interviews. Initially, coding decisions may be quite superficial—by topic, for instance—but later coding decisions are made with the knowledge of, and in consideration of, information gained from all the previously analyzed interviews. Such coding schemes are not superficial, and in light of all the knowledge gained, small pieces of data may have monumental significance. The process is not necessarily superficially objective: It is conducted in light of comprehensive understanding of the significance of each piece of text. The coding process is highly interpretative.

This comprehensive understanding of data bits cannot be acquired in a few objective definitions of each category. Moreover, it cannot be conveyed quickly and in a few definitions to a new member of the research team who has been elected for the purpose of determining a percentage agreement score. This new coder does not have the same knowledge base as the researcher, has not read all the interviews, and therefore does not have the same potential for insight or depth of knowledge required to code meaningfully. Maintaining a simplified coding schedule for the purposes of defining categories for an interrater reliability check will maintain the coding scheme at a superficial level. It will simplify the research to such an extent that all of the richness attained from insight will be lost. Ironically, it forcibly removes each piece of data from the context in which each coding decision should be made. The study will become respectably reliable with an interrater reliability score, but this will be achieved at the cost of losing all the richness and creativity inherent in analysis, ultimately producing a superficial product.

The cost of such an endeavor is equivalent to Mrs. Frisby, who, when the farmer commented that the poisoned rat looked perfectly healthy, said sadly, "Perfectly healthy, but dead!" Your research will be perfectly reliable, but trivial.

There is often a shocked silence when I discuss this with students. But then I ask two questions: "How many of you have written a literature review lately?" Almost every hand is raised. I then ask, "How many of you took a second person to the library with you to make sure you interpreted each article in a manner that was replicable?" Not a single hand remains raised. "Aren't you concerned?" I ask, "How do you know that your analysis, your interpretation of those articles, was reliable?"

The analysis of unstructured, interactive interviews is exactly the same case. Researchers must learn to trust themselves and their judgments and be prepared to defend their interpretations and analyses. But it is death to one's study to simplify one's insights, coding, and analyses so that another person may place the same piece of datum in the same category.

Triangulation Through Distinct Evaluation Teams: The Goal-Free Approach

In program evaluation, an interesting form of team triangulation has been used. Michael Scriven (1972b) has advocated and used two separate teams, one that conducts a traditional goals-based evaluation (assessing the stated outcomes of the program) and a second that undertakes a "goal-free evaluation" in which the evaluators assess clients' needs and program outcomes without focusing on stated goals (see Chapter 4, p. 206). Comparing the results of the goals-based team with those of the goal-free team provides a form of analytical triangulation for determining program effectiveness (Youker & Ingraham, 2014).

Review by Inquiry Participants

Having those who were studied review the findings offers another approach to analytical triangulation. Researchers and evaluators can learn a great deal about the accuracy, completeness, fairness, and perceived validity of their data analysis by having the people described in that analysis react to what is described and concluded. To the extent that participants in the study are unable to relate to and confirm the description and analysis in a qualitative report, questions are raised about the credibility of the findings. In what became a classic study of how evaluations were used, key informants in each case study were asked for both verbal and written reactions to the accuracy and comprehensiveness of the cases. The evaluation report then included those written reactions (Alkin et al., 1979). In her study of homeless youth, Murphy (2014) met with each of the 14 youth to go over the details of the case study she created from their transcribed interviews to affirm accuracy, add additional details and reflections if they so desired, and choose a pseudonym that they wanted to be called in the study, if they had not already done so. (See Thmaris's case study example, pp. 511–516.)

> Obtaining the reactions of respondents to your working drafts is time-consuming, but respondents may (1) verify that you have reflected their perspectives; (2) inform you of sections that, if published, could be problematic for either personal or political reasons; and (3) help you to develop new ideas and interpretations. (Glesne, 1999, p. 152)

Different Purposes Drive Different Review Procedures

Different kinds of studies have different participant review processes, some none at all. Collaborative and participatory inquiry builds in participants' review of cases, quotations, and findings as a matter of course; that's part of what collaboration and participation mean. However, investigative inquiries (Douglas, 1976) aimed at exposing what goes on beyond the public eye are often antagonistic to those in power, so their responses would not typically be used to revise conclusions but might be used to at least offer them an opportunity to provide context and an alternative interpretation. Some traditional social science researchers and evaluators worry that sharing findings with participants for their reactions will undermine the independence of their analysis. Others view it as an important form of triangulation. In an Internet listserv discussion of this issue, one researcher reported this experience:

> I gave both transcripts and a late draft of findings to participants in my study. I wondered what they would object to. I had not promised to alter my conclusions based on their feedback, but I had assured them that my aim was to be sure not to do them harm. My findings included some significant criticisms of their efforts that I feared/expected they might object to. Instead, their review brought forth some new information about initiatives that had not previously been mentioned. And their primary objection was to my not giving the credit for their successes to a wider group in the community. What I learned was not to make assumptions about participants' thinking.

Exhibit 9.3 summarizes three contrasting views of involving those studied in reviewing findings and conclusions.

Critical Friend Review

> A critical friend can be defined as a trusted person who asks provocative questions, provides data to be examined through another lens, and offers critiques of a person's work as a friend. A critical friend takes the time to fully understand the context of the work presented and the outcomes that the person or group is working toward. The friend is an advocate for the success of that work. (Costa & Kallick, 1993, p. 49)

Tessie Tzavaras Catsambasis is president of EnCompass LLC, an international evaluation research company. She is active in evaluation capacity building around the world, including leadership service with the *International Organization for Cooperation in Evaluation*. She also plays the role of critical friend with colleagues' projects and within her own organization. Here's an example she shared with me (and kindly gave permission to include here) that

EXHIBIT 9.3 Different Perspectives on Triangulation by Those Who Were Studied

PERSPECTIVES ON CHECKING IN WITH PARTICIPANTS TO REVIEW THEIR CASES AND OVERALL FINDINGS	RATIONALE
1. Against participants' reviews	a. May raise questions about the independence of findings: risks too much influence by participants on interpretation by the researcher
	b. Not sure what to do if participants and researchers disagree; whose opinion prevails?
2. Weigh pros and cons situationally: It depends	a. Takes time and resources
	b. Must be carefully planned and done well or could create problems in meeting deadlines; may not be worth the hassle
	c. Could be hard to get back to everyone, so some unfairness arises in who gets to review what
	d. Depends on how important and credible it is to the audiences who will receive findings
3. In favor of participant reviews	a. It's the ethical thing to do
	b. It's a chance to correct errors and inaccuracies so you end up with better data
	c. It's a chance to update the data

nicely illustrates the critical friend role as a form of analyst triangulation.

> My team and I conducted an evaluation of a UN organization's Internet-based system that countries could download to track their own HIV/AIDS activities in any sector and area, nationally down to district level. A previous organizational review of the UN organization recommended discontinuing this program based on resource constraints and rumors about problems, but without looking at it closely. The department supporting this program decided to evaluate it first, because they had invested in it significantly and wanted to make a final decision based on evidence. The evaluation we conducted (country visits, focus groups, interviews, survey, benchmarking) revealed many, many problems. But interestingly, some 20 countries were using it (the tracking system). My colleagues who did the data collection were ready to push the button to kill it, citing all the problems we had found. I got involved at the last stage of the data analysis process.
>
> I grilled my colleagues, asking them to justify every conclusion. Their perspective was clear: "This program has so many operational obstacles in the field, no Internet, low capacity, we should recommend discontinuing it." Then, I asked what turned out to be the "turning point" questions: "If this program has so many problems, why are 20 countries choosing to use it?" and "How are those countries addressing the problems you have documented?"
>
> This kind of question (at its best, how is it working?) is an appreciative analytical question asked as a critical friend from a systems dynamics perspective (change the shoes you are wearing, and from the perspective of a country, what do you see?). In response, my colleagues listed many innovations that countries were undertaking to make this tracking program work, and then they concluded, "It is the only option out there that they can control fully, and it is cheap." So "country controlled" was also important, and so was "low cost." Then, I asked them, "Imagine you hold the button to kill the program, do you push it?" They each said, "No, but we would . . ." and proceeded to give me three fabulous recommendations. Their responses enabled us to present to the client the findings, and engage the client in grappling with a tough decision.
>
> Essentially, we said, "This program is filling a demand, and 20 counties are using it in spite of significant

> operational problems. We know you have resource constraints, and this system requires more technical assistance, but if you decide to stop supporting it, consider transferring the system's administration to another funding agency, and also consider certifying independent consultants as technical assistance providers, so countries can contract with them directly for help on the system. And, if you cannot even do that, think about how you will transition in a way that will not hurt countries."
>
> From an evaluation point of view, two things are important: (1) if it were not for these two questions in the analysis, the team would have concluded something very different from the same data, and (2) asking these questions enabled us to facilitate the client to face these challenging findings, have an internal debate about what to do, and own the final decision.

Audience Review as Credibility Triangulation

Reflexive triangulation (Exhibit 2.5, p. 72) includes the audience's reactions to the triangulation mix: (1) the inquirer's reflexive perspective, (2) the perspectives of those studied, and (3) the perspectives of those who received the findings. The opening module of this chapter emphasized that different readers of qualitative reports will apply different criteria to judge quality and credibility. Audience reactions constitute additional data. Whenever possible, I prefer to present draft findings to multiple audiences to learn how they react, what they focus on, what is clear and unclear, and what questions are inadequately answered. In a sense, this is equivalent to theater or movie previews when producers and directors get to gauge audience reaction to a performance or film before it is released. Time and procedures for audience previews and reactions have to be planned in advance, but whenever I've done them, I've been glad I did.

In a study of a community development effort in an inner-city, low-income neighborhood, focus groups were done with diverse groups: African Americans, Native Americans, Hispanics, Hmong residents, and low-income whites. Age-based focus groups were also done: youth under the ages of 25, 25- to 55-year-olds, and those over 55 years. A community advisory group reviewed the study design and voiced no objections to focus groups done homogeneously by either ethnicity or age. In fact, they thought such focus groups were a good idea. But when the draft results were reported in a public meeting that included community people and public officials, the focus group results made it appear that there were great divisions and differences of perspectives among neighborhood ethnic and age-groups. Audience members outside the community were especially focused in on conflicts and differences reported in the findings. Similarities and important areas of agreement got lost amid the reports' overemphasis on differences. Moreover, perspectives within ethnic group and age-group appeared much more monolithic and homogeneous than, in fact, they were. As a result of this feedback, we went back to the field and added heterogeneous focus groups to the data and then drafted a more balanced report. This was in no way undermining inquirer independence. It was making sure we got it right.

Evaluation Audiences and Intended Users

Program evaluation constitutes a particular challenge in establishing credibility because the ultimate test of the credibility of an evaluation report is the response of primary intended users and readers of that report. Their reactions often revolve around *face validity*. On the face of it, is the report believable? Are the data reasonable? Do the results connect to how people understand the world? In seriously soliciting intended users' reactions, the evaluator's perspective is joined to the perspective of the people who must use the findings. Evaluation theorist Ernie House (1977) has suggested that the more "naturalistic" (qualitative) the evaluation, the more it relies on its audiences to reach their own conclusions, draw their own generalizations, and make their own interpretations:

> Unless an evaluation provides an explanation for a particular audience, and enhances the understanding of that audience by the content and form of the argument it presents, it is not an adequate evaluation for that audience, even though the facts on which it is based are verifiable by other procedures. One indicator of the explanatory power of evaluation data is the degree to which the audience is persuaded. Hence, an evaluation may be "true" in the conventional sense but not persuasive to a particular audience for whom it does not serve as an explanation. *In the fullest sense, then, an evaluation is dependent both on the person who makes the evaluative statement and on the person who receives it* [italics added]. (p. 42)

Understanding the interaction and mutuality between the evaluator and the people who use the evaluation, as well as relationships with participants in the program, is critical to understanding the human side of evaluation. This is part of what gives evaluation—and the evaluator—situational and interpersonal "authenticity" (Lincoln & Guba, 1986). Exhibit 9.16, at the end of this chapter (pp. 736–741), provides an experiential account from an evaluator dealing with issues of credibility while building

relationships with program participants and evaluation users; her reflections provide a personal, in-depth description of what *authenticity* is like from the perspective of one participant-observer.

Expert Audit Review

A final review alternative involves using experts to assess the quality of analysis or, where the stakes for external credibility are especially high, performing a meta-evaluation or process audit. An external audit by a disinterested expert can render judgment about the quality of data collection and analysis. "That part of the audit that examines the process results in a *dependability judgment* [italics added], while that part concerned with the product (data and reconstructions) results in a *confirmability judgment* [italics added]" (Lincoln & Guba, 1986, p. 77). Such an audit would need to be conducted according to appropriate criteria. For example, it would not be fair to audit an aesthetic and evocative qualitative presentation by traditional social science standards or vice versa. But within a particular framework, expert reviews can increase credibility for those who are unsure how to distinguish high-quality work. That, of course, is the role of the doctoral committee for graduate students and peer reviewers for scholarly journals. Problems arise when peer reviewers apply traditional scientific criteria to constructivist studies, and vice versa. In such cases, the review or audit itself lacks credibility. Exhibit 9.4 on the next page presents an example of an expert meta-evaluation (evaluation of the evaluation) to independently judge the quality and establish credibility for a high-stakes international mixed-methods evaluation.

The challenge of getting the right expert, one who can apply an appropriately critical eye, is wittily illustrated by a story about the great French artist Pablo Picasso. Marketing of fakes of his paintings plagued Picasso. His friends became involved in helping check out the authenticity of supposed genuine originals. One friend in particular became obsessed with tracking down frauds and brought several paintings to Picasso, all of which the master identified as fake. A poor artist who had hoped to profit from having obtained a Picasso before the great artist's works had become so valuable sent his painting for inspection via the friend. Again Picasso pronounced it a forgery.

> "But I saw you paint this one with my very own eyes," protested the friend.
>
> "I can paint false Picassos as well as anyone," retorted Picasso.

4. Theory Triangulation

> Greek legend tells of the fearsome hotelier Procrustes who would adjust his guests to match the length of his bed, stretching the short and trimming off the legs of the tall. Guides to program theory that are too prescriptive risk creating such a Procrustean bed. When the same approach to program theory is used for all types of interventions and all types of purposes, the risk is that the interventions will be distorted to fit into a preconceived format. Important aspects may be chopped off and ignored, and other aspects may be stretched to fit into preconceived boxes of a factory model, with inputs, processes, outcomes, and impacts.
>
> Purposeful program theory requires thoughtful assessment of circumstances, asking in particular, "Who is going to use the program theory and for what purposes?" and "What is the nature of the intervention and the situation in which it is implemented?" It requires a wide repertoire, not a one-size-fits-all approach to program theory.
>
> Purposeful program theory also requires attention to the limitations of any one program theory, which must necessarily be a simplification of reality and a willingness to revise it as needed to address emerging issues.
>
> —Funnell and Rogers (2011, p. xxi)
> ***Purposeful Program Theory***

Having discussed triangulation of qualitative data sources, mixed-methods triangulation, and multiple analyst triangulation, we turn now to the fourth and final kind of triangulation: using different theoretical perspectives to look at the same data.

EXHIBIT 9.4 Metaevaluation: Evaluating the Evaluation of the Paris Declaration on Development Aid

It has become a standard in major high-stakes evaluations to commission an independent review to determine whether the evaluation meets generally accepted standards of quality and, in so doing, to identify strengths, weaknesses, and lessons (Stufflebeam & Shrinkfield, 2007, p. 649). The major addition to the Joint Committee Standards for Evaluation, when revised in 2010, was that of "Evaluation Accountability Standards" focused on meta-evaluation.

Evaluation Accountability Standards

> ***E1 Evaluation documentation:*** Evaluations should fully document their negotiated purposes and implemented designs, procedures, data, and outcomes.
>
> ***E2 Internal meta-evaluation:*** Evaluators should use these and other applicable standards to examine the accountability of the evaluation design, procedures employed, information collected, and outcomes.
>
> ***E3 External meta-evaluation:*** Program evaluation sponsors, clients, evaluators, and other stakeholders should encourage the conduct of external meta-evaluations using these and other applicable standards (Joint Committee on Standards, 2010; Yarbrough, Shulha, Hopson, & Caruthers, 2010).

Evaluating the Evaluation of the Paris Declaration

Given the historic importance of the Evaluation of the Paris Declaration on Development Aid (Dabelstein & Patton, 2013b), the Management Group overseeing the evaluation commissioned an independent assessment of the evaluation. Prior to undertaking this review, we had no prior relationship with any members of the Management Group or the Core Evaluation Team. We had complete and unfettered access to any and all evaluation documents and data, and to all members of the International Reference Group, the Management group, the Secretariat, and the Core Evaluation Team. Our evaluation of the evaluation included reviewing data collection instruments, templates, and processes; reviewing the partner country and donor evaluation reports on which the synthesis of findings was based; directly observing two meetings of the International Reference Group where the evidence was examined and the conclusions refined and sharpened accordingly; engaging International Reference Group participants in a reflective practice, lessons-learned session; surveying participants about the evaluation process and partner country evaluations; and interviewing key people involved in and knowledgeable about how the evaluation was conducted. The evaluation of the evaluation included assessing both the evaluation report's findings and the technical appendix that details how the findings were generated. The Development Assistance Committee (DAC) of the Organization for Economic Co-operation and Development (OECD) established international standards for evaluation in 2010, and those were the standards used for the meta-evaluation (OECD-DAC, 2010). A meta-evaluation audit statement confirming the quality, credibility, and usability of the evaluation was included as a preface to the full evaluation reports. The meta-evaluation report (Patton & Gornick, 2011a) was published and made available online two weeks after the Final Evaluation report was published. This timing was possible because the meta-evaluation began halfway through the Paris Declaration Evaluation and the meta-evaluation team had access to draft versions of the final report at each stage of the report's development. The process for conducting the meta-evaluation and its uses are discussed in detail in Patton (2013).

The Paris Declaration Evaluation received the 2012 American Evaluation Association (AEA) Outstanding Evaluation Award. At the award ceremony, the chair of the AEA Awards Committee, Frances Lawrenz (2013), summarized the merits of the evaluation that led to the award selection and recognition:

> The success of the Paris Declaration Phase 2 Evaluation required an unusually skilled, knowledgeable and committed evaluation team; a visionary, well-organized, and well-connected Secretariat to manage the logistics, international stakeholder meetings, and financial accounts; and a highly competent and respected Management Group to provide oversight and ensure the Evaluation's independence and integrity. This was an extraordinary partnership where all involved understood their roles, carried out their responsibilities fully and effectively, and respected the contributions of other members of the collaboration.

Chapter 3 presented a number of general theoretical frameworks derived from diverse intellectual and disciplinary traditions. More concretely, multiple theoretical perspectives can be brought to bear on specialized substantive issues. For example, one might examine interviews with therapy clients from different psychological perspectives: psychotherapy, Gestalt, Adlerian, and behavioral psychology. Observations of a group, community, or organization can be examined from a Marxian or Weberian perspective, a conflict or functionalist point of view. The point of theory triangulation is to understand how differing assumptions and premises affect findings and interpretations.

"I envy your confidence. Even after decades of evaluations, these metaevalutions still make me feel naked."

Examples of Theory Triangulation

Let's suppose we are studying famine in a drought-afflicted region of an African country. We have quantitative data on food production (sorghum and millet), nutrition data from household surveys, health data from clinics, rainfall data over many years, interviews with villagers (males and females), key informant interviews (e.g., government officials, agricultural experts, aid agency staff members, and village leaders), and case studies of purposefully sampled villages telling the story of their agricultural and nutritional situations and experiences before and during the famine. Put all of these data together and we have an in-depth description of the extent and nature of the famine, its effects on subsistence agriculture families, food and agricultural assistance provided, and the interventions of government and international agencies. We have (a) mixed-methods triangulation and (b) multiple sources of qualitative data (interviews, observations, case studies, documents), and (c) our team members have analyzed the patterns independently to confirm the findings as well as had the findings externally reviewed by experts. Thus, we can make a credible case for the nature, extent, and impacts of the famine. What does theory triangulation add?

When we move from description to interpretation, we need a framework to make sense of and explain the patterns in the data. Why is the region experiencing famine? Why aren't interventions more effective? Different theoretical frameworks emphasize different explanatory variables.

- Climate change theory would emphasize long-term weather and climate trends.
- Malthusian theory would emphasize overpopulation.
- Marxian theory would emphasize power dynamics (Who controls the means of production? How do the powerful benefit from famine?).
- Weberian theory would emphasize organizational competence and incompetence (How does the functioning and activities of government and international agencies exacerbate or alleviate famine?).
- Ecological systems theory would call for examining the interactions between the ecosystem, farming practices, soil and water conditions, and markets.
- Cultural systems theory would emphasize the way in which cultural beliefs and norms affect the experience of and responses to famine by the people affected.
- Feminist theory would point to the role of women in the system as a factor in how the famine affects families and their responses to the crisis (Podems, 2014b).
- Cognitive theory would focus on how people make decisions in the face of changing conditions.

When designing the famine study, these various theoretical perspectives would inform the kinds of questions to be asked and data to be collected. When analyzing the findings and explaining results, these diverse theoretical perspectives provide competing interpretations for explaining the patterns and observed impacts. *Theory triangulation* involves examining the data through different theoretical lenses to see what theoretical framework (or combination) aligns most convincingly with the data (best fit).

Theory triangulation for evaluation can involve examining the data from the perspectives of various stakeholder positions. It is common for diverse stakeholders to disagree about program purposes, goals,

and means of attaining goals. These differences represent different "theories of action" (Patton, 2012a) that can cast the same findings in different perspective-based lights. When we were seeking explanations to explain dropout rates for adult literacy programs in Minnesota, the predominant staff theory was that low-income people led chaotic lives and couldn't manage regular attendance and follow-through in a program. Political explanations included laziness, effects of multigenerational poverty, lack of good jobs to motivate participants to complete programs, and cultural deprivation theories. But the explanation that best fit the data (interviews with dropouts) was that the adult literacy programs were lousy learning experiences: large class sizes; disinterested and disrespectful teachers, poorly paid and exhausted from having already taught all day in their regular jobs; uninteresting and outdated curriculum materials; and an all-around depressing environment. Most traditional explanations blamed the participants or the larger societal problems that affected the participants, but the actual data pointed to ineffective programs, something that was actionable. Changes were made, and dropout rates went down significantly.

SIDEBAR

THEORY INTEGRATION MEETS THEORY TRIANGULATION

Different criteria for evaluating the quality of qualitative remain fluid as qualitative inquirers move back and forth among genres, ignoring the boundaries, much as birds ignore human fences—except to use them occasionally as convenient places to rest. Consider the reflections on working across and integrating multiple genres and theoretical orientations of self-described "critical educators" Patricia Burdell and Beth Blue Swadener (1999). They combine autobiographical narratives with a variety of theoretical perspectives, including critical, dialogic, phenomenological, feminist, and semiotic perspectives. They speculate that "it is perhaps both the intent and effect of many of these texts to broaden the 'acceptable' or give voice to the intellectual contradictions and tensions in everyday lives of scholar-teachers and researchers" (p. 23).

> Our research has used narrative inquiry, collaborative ethnography, and applied semiotics. Between us, we share an identity and scholarship in critical and feminist curriculum theory. We are frequent border-crossers. We seek texts that allow us to enter the world of others in ways that have us more present in their experience, while better understanding our own. (p. 23).

They call this border-crossing genre "critical personal narrative and autoethnography." The real world in which inquiry occurs is not a very neat and orderly place. Nor is it likely to become so. Theoretical and methodological border crossers are natural and determined *triangulators.*

Thoughtful, Systematic Triangulation

All four of these different types of triangulation—(1) mixed-methods triangulation, (2) triangulation of qualitative data sources, (3) analyst triangulation, and (4) theory or perspective triangulation—offer strategies for reducing systematic bias and distortion during data analysis, and thereby increasing credibility. In each case, the strategy involves checking findings against other sources and perspectives. Triangulation, in whatever form, increases credibility and quality by countering the concern (or accusation) that a study's findings are simply an artifact of a single method, a single source, or a single investigator's blinders. Exhibit 9.1 (p. 660) reviews and summarizes the four types of triangulation (items 7–10 in Exhibit 9.1).

Exhibit 9.5 presents a model for rigorous analysis that broadens and deepens triangulation processes in high-stakes, high-visibility situations. Eight attributes of a rigorous analysis process were identified by studying experienced intelligence analysts from multiple U.S. federal investigative agencies. The researchers used a cognitive systems approach in which professional intelligence analysts were engaged in going beyond assessment of the quality of an analysis based on product quality to examine the analytic processes necessary to generate a high-quality, credible, and useful product. The understanding of rigor that emerged was that it is not about following a standardized, highly prescribed analytical process (a formula or recipe) but, rather, "assessing the contextual sufficiency of many different aspects of the analytic process" (Zelik et al., 2007). The researchers posited that these dimensions could be relevant to any process where analysts must make sense of complex data, and the rigor and resulting credibility of their analytical process will affect the utility of the findings for decision making. Examine Exhibit 9.5 carefully and thoughtfully. There's a lot there pulled together in a comprehensive, coherent, and integrated triangulation model: The Rigor Attribute Model. What comes across most powerfully from the work that generated the model is that a product (report, findings, or presentation of results) cannot be assessed for quality and credibility without knowing the nature and rigor of the analytical process that generated the findings. That insight is consistent with the focus of my MQP Rumination, avoiding research rigor mortis, in this chapter (see pp. 701–703).

EXHIBIT 9.5 Dimensions of Rigorous Analysis and Critical Thinking

The Rigor Attribute Model

Eight attributes of a rigorous analysis process were identified by studying experienced intelligence analysts from multiple U.S. federal investigative agencies. The researchers used a cognitive systems approach in which professional intelligence analysts were engaged in going beyond assessment of the quality of an analysis based on product quality to examine the analytic process that generated the product. The understanding of rigor that emerged was that it is not about following a standardized process but, rather, "assessing the contextual sufficiency of many different aspects of the analytic process" (Zelik et al., 2007). The researchers posited that these dimensions could be relevant to any process where analysts must make sense of complex data, and the rigor and resulting credibility of their analytical process will affect the utility of the findings for decision making.

Overview of the Eight Dimensions of Rigorous Analysis

RIGOR ATTRIBUTE	HIGH-RIGOR PROCESS	LOW-RIGOR PROCESS
1. *Hypothesis exploration*. Extent to which multiple hypotheses were seriously examined against the data	Test multiple hypotheses to identify the best, most probable explanations	Minimal weighing of alternatives
2. *Information search*. Depth and breadth of the search process used in collecting data	Comprehensively explore as much data as relevant to the inquiry (diligent, purposeful sampling)	Data collection limited to routine and readily available data sources (convenience sampling)
3. *Information validation*. Information sources are corroborated and cross-validated (triangulation)	Systematic verification and triangulation of information and sampling information-rich, trustworthy, and knowledgeable sources	Little effort made to triangulate (use converging evidence to verify source accuracy)
4. *Stance analysis*. Evaluation of data to identify and contextualize the perspective of the source	Investigate key informants' backgrounds to assess how their perspective might bias information they provide	Nothing is done when clear bias in a source is detected
5. *Sensitivity analysis*. The extent to which analysts consider, understand, and make explicit the assumptions, strengths, weaknesses, limitations, and gaps of their analysis	Systematic and strategic assessment of implications for interpretations, conclusions, and explanations if elements of the supporting sources and evidence prove invalid, inadequate, or otherwise problematic	Explanations accepted and reported if they seem appropriate and valid on a surface level. Emphasis on face validity
6. *Specialist collaboration*. The degree to which an analyst incorporates the perspectives of experts and key knowledgeables into their assessments	The analyst has talked to, or may be, a leading expert in the key content areas of the analysis. Seeks out independent, external expert peer review for high-stakes analysis	Little or no effort is made to seek out and incorporate independent, external expertise; no peer review before reporting
7. *Information synthesis*. Refers to how far beyond simply collecting, listing, and analyzing distinct data elements, sources, and cases an analyst went in the interpretive process	Extracted and integrated information with a thorough consideration of diverse interpretations of relevant data, noting both areas of consistency in findings and areas where different methods and data yield conflicting findings	Analyst simply complies and reports the relevant information in a sequential and compartmentalized form; little or no integration and synthesis

(Continued)

(Continued)

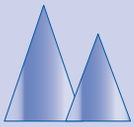
RIGOR ATTRIBUTE	HIGH-RIGOR PROCESS	LOW-RIGOR PROCESS
8. *Explanation critique.* A form of collaboration that engages different perspectives in examining the preponderance of evidence supporting primary conclusions	Peers and experts are involved in independently examining the interpretive chain of reasoning and inferences made, explicitly distinguishing which are stronger and which weaker	Little or no use of other analysts to give input on explanation quality

SOURCE: Adapted and revised from Zelik, Patterson, and Woods (2007).

SIDEBAR

INTERPRETING TRIANGULATION RESULTS: MAKING SENSE OF CONFLICTING AND INCONSISTENT CONCLUSIONS

A common misconception about triangulation involves thinking that the purpose is to demonstrate that different data sources or inquiry approaches yield essentially the same result. The point is to test for such consistency. Different kinds of data may yield somewhat different results because different types of inquiry are sensitive to different real-world nuances. Different theoretical frameworks will likely foster different interpretations of the same findings. Different analysts may well interpret the same patterns in different ways. Thus, *understanding inconsistencies in findings across different kinds of triangulation can be illuminative and important.* Finding such inconsistencies ought not to be viewed as weakening the credibility of results but, rather, as offering opportunities for deeper insight into the relationship between inquiry approach and the phenomenon under study.



Alternative and Competing Criteria for Judging the Quality of Qualitative Inquiries, Part 1

Universal Criteria, and Traditional Scientific Research Versus Constructivist Criteria

> Every way of seeing is also a way of not seeing.
>
> —David Silverman (2000, p. 825)

Judging Quality: The Necessity of Determining Criteria

It all depends on criteria. Judging quality requires criteria. Credibility flows from those judgments. Quality and credibility are connected in that judgments of quality constitute the foundation for perceptions of credibility.

Diverse approaches to qualitative inquiry—phenomenology, ethnomethodology, ethnography, hermeneutics, symbolic interaction, heuristics, critical theory, realism, grounded theory, and feminist inquiry, to name but a few—remind us that issues of quality and credibility intersect with audience and intended inquiry purposes. Research directed to an audience of independent feminist scholars, for example, may be judged by somewhat different criteria from research addressed to an audience of government economic policymakers. Formative research or action inquiry for program improvement involves different purposes and therefore different criteria of quality compared with summative evaluation aimed at making fundamental continuation decisions about a program or policy. Thus, it is important to acknowledge at the outset that particular philosophical underpinnings or theoretical orientations and special purposes for qualitative inquiry will generate different criteria for judging quality and credibility.

Despite this, efforts to generate universal criteria and checklists for quality abound. The results are as follows: multiple possibilities, no consensus, and ongoing debate.

A Review of Quality Assurance Recommendations for Qualitative Research

An interdisciplinary team of health researchers engaged in worldwide malaria prevention and treatment set out to identify quality criteria for qualitative research and evaluation (Reynolds et al., 2011). They found 93 papers published between 1994 and 2010 that offered and discussed quality criteria, 37 of which were sufficiently detailed to merit further analysis. (The 56 papers that were rejected focused only on review criteria for publication or guidance on a specific qualitative method or single stage of the research process, such as data analysis.) They found no consensus about how to ensure the quality of qualitative research. However, they were able to categorize approaches into two "narratives" about quality: (1) an output-oriented approach versus (2) a process-oriented approach:

1. The most dominant narrative detected was that of an *output-oriented approach*. Within this narrative, quality is conceptualized in relation to theoretical constructs such as validity or rigor, derived from the positivist paradigm, and is demonstrated by the inclusion of certain recommended methodological techniques: the use of triangulation, member (or participant) validation of findings, peer review of findings, deviant or negative case analysis, and multiple coders of data.

Strengths of the output-oriented approach for assuring quality of qualitative studies include the acceptability and credibility of this approach within the dominant positivist environment where decision making is based on "objective" criteria of quality. Checklists equip those unfamiliar with qualitative research with the means to assess its quality.

The weakness of this approach is that "following of check-lists does not equate with understanding of and commitment to the theoretical underpinnings of qualitative paradigms or what constitutes quality within the approach. The privileging of guidelines as a mechanism to demonstrate quality can mislead inexperienced qualitative researchers as to what constitutes good qualitative research. This runs the risk of reducing qualitative research to a limited set of methods, requiring little theoretical expertise and diverting attention away from the analytic content of research unique to the qualitative approach. Ultimately, one can argue that a solely output-oriented approach risks the values of qualitative research becoming skewed towards the demands of the positivist paradigm without retaining quality in the substance of the research process."

2. By contrast, the second, *process-oriented* narrative, presented conceptualizations of quality that were linked to principles or values considered inherent to the qualitative approach, to be understood and enacted throughout the research process. Six common principles were identified across the narrative: (1) reflexivity of the researcher's position, assumptions, and practice; (2) transparency of decisions made and assumptions held; (3) comprehensiveness of approach to the research question; (4) responsibility toward decision making acknowledged by the researcher; (5) upholding good ethical practice throughout the research; and (6) a systematic approach to designing, conducting, and analyzing a study.

> Strengths of the process-oriented approach include the ability of the researcher to address the quality of their research in relation to the core principles or values of qualitative research. The core principles identified in this narrative also represent continuous, researcher-led activities rather than externally determined indicators such as validity, or endpoints. Reflexivity, for example, is an active, iterative process—***an attitude of attending systematically to the context of knowledge construction . . . at every step of the research process.*** As such, this approach emphasises the need to consider quality throughout the whole course of research, and locates the responsibility for enacting good qualitative research practice firmly in the lap of the researcher(s).

Need for a Flexible Quality Framework

The review team (Reynolds et al., 2011) found that "there is an increasing demand for the qualitative research field to move forward in developing and establishing coherent mechanisms for quality assurance of qualitative research." They concluded with a recommendation for "the development of a flexible framework to help qualitative researchers to define, apply and demonstrate principles of quality in their research." They further recommended that "the strengths of both the output-oriented and process-oriented narratives be brought together to create guidance that reflects core principles of qualitative research but also responds to expectations of the global health field for explicitly assured quality in research."

> We recommend the development of a framework that helps researchers identify their core principles, appropriate for their epistemological and methodological approach, and ways to demonstrate that these have been upheld throughout the research process. . . . We propose that this framework be flexible enough to accommodate different qualitative methodologies without dictating essential activities for promoting quality. (Reynolds et al., 2011)

This chapter addresses this recommendation, offering both a generic quality framework as well as specialized quality criteria for specific types of qualitative inquiry.

SIDEBAR

THE PURPOSE OF AND DEBATE ABOUT CRITERIA

> Criteria are standards, benchmarks, norms, and, in some cases, regulative ideals that guide judgments about the goodness, quality, validity, truthfulness, and so forth of competing claims (or methodologies, theories, interpretations, etc.). . . .
>
> Criteria that have been proposed for judging the processes and products of social inquiry include truth, relevance, validity, credibility, plausibility, generalizability, social action, and social transformation, among others. Some of these criteria are epistemic (i.e., concerned with justifying knowledge claims as true, accurate, correct), others are political (i.e., concerned with warranting the power, use, and effects of knowledge claims or the inquiry process more generally); still others are moral or ethical standards (i.e., concerned with the right conduct of the inquirer and the inquiry process in general). . . .
>
> Poststructuralist and postmodernist approaches to qualitative inquiry are also shaping the way we conceive of criteria. Given the growing influence of narrative approaches and experimental texts in qualitative inquiry, it is becoming more common to find discussions of rhetorical and aesthetic criteria replacing discussions of epistemic criteria. Other scholars argue that epistemological criteria cannot be neatly decoupled from political and critical agendas and ethical concerns. Some scholars in qualitative inquiry have little patience for discussing criteria within different epistemological frameworks and theoretical perspectives and prefer to focus on the craft of using various methodological procedures for producing "quality" work.
>
> —Schwandt (2007, pp. 49–50)
> *The Sage Dictionary of Qualitative Inquiry*

Judging the Quality of Alternative Approaches to Qualitative Inquiry

There can be no universal, generic, standardized, and all-encompassing criteria for judging the quality of qualitative studies because qualitative inquiry is not monolithic, uniform, or standardized. It's as if someone set out to create a universal checklist for beauty that ignored culture, human variability, variety, and differences in taste, socialization, and values (oh yes, the "Miss Universe" and "Miss World" contests notwithstanding). The common core elements across all kinds of qualitative inquiry are attention to language, words, narrative, description, stories, cases, worldviews, and how people make sense of their worlds. Tracy (2010), for example, identified "eight 'big-tent' criteria for excellent qualitative research": (1) worthy topic, (2) rich rigor, (3) sincerity, (4) credibility, (5) resonance, (6) significant contribution, (7) ethics, and (8) meaningful coherence. But approaches to inquiring into and judging attainment of these general criteria are diverse and multifaceted, serve competing purposes, and are, ultimately, a matter of debate (Gordon & Patterson, 2013).

It is possible to specify quality criteria for research generally. These are not unique to qualitative inquiry but apply to scientific inquiries of all kinds. Exhibit 9.6 presents these general, science-based quality criteria.

EXHIBIT 9.6 General Scientific Research Quality Criteria

QUALITY CRITERIA	ELABORATION/EXAMPLES
1. Clarity of purpose	Basic research, applied research, and evaluation research, for example, serve different purposes and are judged by different standards. (See Exhibit 5.1, p. 250.)
2. Epistemological clarity	Inquiry traditions like positivism, naturalism, social construction, realism, pheneomenology, and pragmatism are based on different criteria about what constitutes knowledge, how it is acquired, and how it should be judged. (See Chapter, especially Exhibit 3.3, pp. 97–99.)
3. Questions and hypotheses flow from and are consistent with purpose and epistemology	Different purposes and iquiry traditions emphasize different priority questions. (See Exhibit 3.3, pp. 97–99.)
4. Methods, design, and data collection procedures are aprorpriate for the nature of the inquiry	Purpose, epistemology, and research questions, in combination, drive methods, design, and data collection decisions. Matching methods to questions and hypotheses, given constraints of time, resources, and access, is basic.
5. Data collection procedures are systematic and carefully documented	The foundation of all science is careful methodological documentation so that those reviewing findings can determine how they were produced.
6. Data analysis is appropriate for the kind of data collected	Matching analytical procedures to the nature and type of data collected is a basic standard. There may be disagreemnts about what is appropriate, but the researcher has the obligation to make the case for appropriateness and justify methodological and analytical decsions made.
7. Strengths and weaknesses are acknowledged and discussed	No studies are perfect. All have limitations. These should be acknowledged and their implications for interpreting findings discussed.
8. Findings should flow from the data and analysis	The connection between data collected, analysis undertaken, and findings (conclusions, explanations) should be clear and explained.
9. Research should be presented for review	A fundamental principle of science is openness to review by those in a position to judge quality.
10. Ethical reflection and disclosure	All scientific traditions and disciplines have ethical standards, like avoiding (or at least disclosing) conflicts of interest and treating human subjects with respect. Compliance with ethical standards should be discussed.

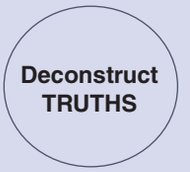
EXHIBIT 9.7 Alternative Sets of Criteria for Judging the Quality and Credibility of Qualitative Inquiry

1. **Traditional Scientific Research Criteria**

   1. Objectivity of the inquirer (minimize bias)
   2. Hypothesis generation and testing
   3. Validity of the data
   4. Interrater reliability of codings and pattern analyses
   5. Conclusions about the correspondence of findings to reality
   6. Generalizability (external validity)
   7. Strength of causal explanations (attribution analysis)
   8. Contributions to theory
   9. Independence of conclusions and judgments
   10. Credibility to knowledgeable disciplinary researchers (peer review)

(These criteria are explained and discussed on pp. 683–684.)

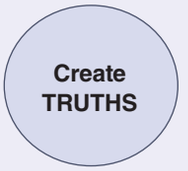
2. **Social Construction and Constructivist Criteria**



   1. Subjectivity acknowledged (discuss and take into account inquirer perspective)
   2. Trustworthiness and authenticity
   3. Interdependence: relationship based (intersubjectivity)
   4. Triangulation (capturing and respecting multiple perspectives)
   5. Reflexivity
   6. Particularity (doing justice to the integrity of unique cases)
   7. Enhanced and deepened understanding (*verstehen*)
   8. Contributions to dialogue
   9. Extrapolation and transferability
   10. Credible to and deemed accurate by those who have shared their stories and perspectives

(These criteria are explained and discussed on pp. 684–686.)

3. **Artistic and Evocative Criteria**



   1. Emotionally evocative: connects with and moves the audience
   2. Integrates science and art to open the world to us
   3. Creativity
   4. Aesthetic quality, artistic representation
   5. Interpretive vitality, sensuous
   6. Embedded in lived experience
   7. Stimulating and provocative
   8. Voice distinct, expressive
   9. *Feels* "true" or "authentic" or "real"
   10. Crystallization

(These criteria are explained and discussed on pp. 687–690.)

4. **Participatory and Collaborative Criteria**

   1. Genuine and significant participation from inquiry focus, through design, data collection, analysis, and reporting; participation is real
   2. Researchers and participants are co-inquirers, sharing power and decision making
   3. Interactive validity and interpersonal competence

4. Builds capacity through learning by doing
5. Mutual respect
6. Group reflexivity
7. Interdependence
8. Sense of group ownership ("We did this.")
9. Group accountability: negotiated trade-offs explicit and transparent
10. Credibility within the group the basis for external credibility

(These criteria are explained and discussed on pp. 690–691.)

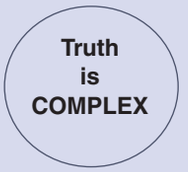
5. **Critical Change Criteria**

1. Critical perspective: increases consciousness about injustices
2. Identifies nature and sources of inequalities and injustices
3. Represents the perspective of the less powerful
4. Makes visible the ways in which those with more power exercise and benefit from power
5. Engages those with less power respectfully and collaboratively
6. Builds the capacity of those involved to take action
7. Identifies potential change-making strategies
8. Praxis
9. Clear historical and values context
10. Consequential validity

(These criteria are explained and discussed on pp. 691–693.)

6. **Systems Thinking and Complexity Criteria**

1. Analyze and map systems of interests
2. Attend to interrelationships
3. Capture perspectives
4. Sensitive to and explicit about boundary implications
5. Capture emergence
6. Expect and document nonlinearities
7. Adapt inquiry in the face of uncertainties
8. Describe systems changes and their implications
9. Contribution analysis
10. Credible to systems thinkers



(These criteria are explained and discussed on pp. 693–695.)

7. **Pragmatic, Utilization-Focused Criteria**

1. Focus inquiry on informing action and decisions
2. Identify intended uses and users
3. Interactive engagement with intended users to enhance relevance and use
4. Practical orientation throughout
5. Relevance to real-world issues and concerns
6. Time findings and feedback to support use
7. Understandable methods and findings
8. Actionable findings
9. Credible to primary intended users
10. What is useful is true
11. Extract lessons

(These criteria are explained and discussed on pp. 695–697.)

From the General to the Particular: Seven Sets of Criteria for Judging the Quality of Different Approaches to Qualitative Inquiry

Once we move beyond general criteria for scientific inquiry (Exhibit 9.6) to address specific quality criteria for qualitative inquiry, we must move from the general to the particular and contextual. Exhibit 9.7 lists criteria that are embedded in and flow from distinct qualitative inquiry frameworks. The *traditional scientific research criteria* are embedded in and derived from what I discussed in Chapter 3 as *reality-testing inquiry* frameworks that include positivist, postpositivist, empiricist, and foundationalist epistemologies pp. 105–108. The *social construction criteria* are derived from the discussion of "constructivism" in Chapter 3 pp. 121–126. The *artistic and evocative criteria* are derived from the discussion of autoethnography and evocative forms of inquiry in Chapter 3, especially the criteria suggested by Richardson (2000b) for "creative analytic practice of ethnography." The fourth set of criteria, participatory and collaborative approaches, are based on traditions and approaches reviewed in Chapter 4 pp. 213–222. The fifth set of criteria, *critical change criteria*, flow from critical theory, feminist inquiry, activist research, and participatory research processes aimed at empowerment. The sixth set of criteria, *systems and complexity criteria*, are derived from the discussion in Chapter 3 pp. 139–151.The seventh and final set of criteria, *pragmatic and utilization-focused criteria*, are based on discussions in Chapters 3 and 4 pp. 152–157 as well as program evaluation standards and principles (Joint Committee and Standards, 2010) and "Guiding Principles for Evaluators" (AEA Task Force on Guiding Principles for Evaluators, 1995).

To some extent, all of the theoretical, philosophical, and applied orientations reviewed in Chapters 3 and 4 provide somewhat distinct criteria, or at least priorities and emphases, for what constitutes a quality contribution within those particular perspectives and concerns. I've chosen these seven broader sets of criteria to capture the primary debates that differentiate qualitative approaches and, more specifically, to highlight what seem to me to differentiate *reactions* to qualitative inquiry. In this chapter, we are primarily concerned with how others respond to our work. With what perspectives and by what criteria will our work be judged by those who encounter and engage it?

Some of the confusion that people have in assessing qualitative research stems from thinking it represents a uniform perspective, especially in contrast to quantitative research. This makes it hard for them to make sense of the competing approaches within qualitative inquiry. By understanding the criteria that others bring to bear on our work, we can anticipate their reactions and help them position our intentions and criteria in relation to their own expectations and criteria. In terms of the Reflexive Triangulated Inquiry model presented in Chapter 2 as Exhibit 2.5 (see p. 72), we're dealing here with the intersection between the inquirer's perspective and the perspective of those receiving the study (the audiences).

SIDEBAR

DIFFERENT AUDIENCES INTERESTED IN AND INVOLVED IN ASSESSING THE QUALITY OF QUALITATIVE RESEARCH AND EVALUATION

Criteria of quality can and often do vary by audience (Flick, 2007b, pp. 3–8). Here are some questions to consider in thinking about the intersection of quality criteria and audience.

1. *Your criteria.* You, the inquirer, presumably have an interest in doing quality work. How do you decide what standards and criteria of quality you will adhere to?
2. *Primary users of your findings.* Others will read and potentially use your findings. Who are the intended users of what you generate, and what criteria will they apply in judging the quality and credibility of your work?
3. *Funders of your inquiry.* If your inquiry has been funded by a grant, an agency, an evaluation contract, or some other funding mechanism, funders will be judging whether what you produced was worth what it cost. How will they make that judgment?
4. *Publication reviewers.* You may want to publish your findings. How will journals, book editors, and peer reviewers judge your work?

You need not be passive about others' criteria and judgments. Indeed, you ought not to be passive. You should make explicit the quality criteria you have applied in designing and implementing your inquiry and invite readers, funders, and peer reviewers to join you in using your criteria. You may also add the caveat that if they apply different criteria, their judgments of quality may well differ from yours and from those who follow the criteria you're operating under. In all of this keep in mind that

> the question of how to ascertain the quality of qualitative research has been asked since the beginning of qualitative research and attracts continuous and repeated attention. However, answers to this question have not been found—at least not in a way that is generally agreed upon. (Flick, 2007b, p. 11)

Criteria Determine What We See: The Umpires' Perspectives

Different perspectives about things such as truth and the nature of reality constitute paradigms or worldviews based on alternative epistemologies and ontologies. People viewing qualitative findings through different paradigmatic lenses will react differently just as we, as researchers and evaluators, vary in how we think about what we do when we study the world. These differences are nicely illustrated by the classic story of three baseball umpires who, having retired after a game to a local establishment for the dispensing of reality-distorting but truth-enhancing libations, are discussing how they call balls and strikes.

> "I call them as I see them," says the first.
>
> "I call them as they are," says the second.
>
> "They ain't nothing until I call them," says the third.

That's the classic version of the story. Now, thanks to high-speed camera technology, we can update the story.

As chance would have it, two management researchers, Brayden King of Northwestern University and Jerry Kim of Columbia Business School, happened to be in the same bar going over their research on the accuracy of umpires' calls. Overhearing the three umpires, they went up to them and said, "Fourteen percent of the time you call them wrong." Before the umpires could argue, they explained,

> We analyzed more than 700,000 pitches thrown during the 2008 and 2009 seasons. In addition to an average error rate of 14%, we found that umpires tended to favor the home team and that umpires were more likely to make mistakes when the game was on the line. (Based on King & Kim, 2014, p. SR12)

The two researchers went on like this for some time, breaking down the error rates by innings, situation, pitcher and batter ethnicity and race, pitcher reputation, and so forth and so on, until finally the umpires together put up their hands and told them to stop.

> The first umpire said, "Your criteria are based on a high-speed camera. We get that. You love your numbers. We get that. And your analysis is interesting, even fascinating. We get that. But during a game we don't use a camera. So I still call them as I see them," he reiterated.
>
> "And I call them as they are," repeated the second.
>
> "And they ain't nothing until I call them," concluded the third.

We turn now to discussion and elaboration of the seven alternative sets of criteria for judging the quality of qualitative work summarized in Exhibit 9.7.

1. Traditional Scientific Research Criteria

> **The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.**
>
> —Isaac Asimov (1920–1992)
> Science author and science fiction writer

One way to increase the credibility and legitimacy of qualitative inquiry among those who place priority on traditional scientific research criteria is to emphasize those criteria that have priority within that tradition. Science has traditionally emphasized objectivity, so qualitative inquiry within this tradition emphasizes procedures for minimizing investigator bias. Those working within this tradition will emphasize rigorous and systematic data collection procedures, for example, cross-checking and cross-validating sources during fieldwork. In analysis it means, whenever possible, using multiple coders and calculating intercoder consistency to establish the validity and reliability of pattern and theme analysis. Qualitative researchers working in this tradition are comfortable using the language of "variables" and "hypothesis testing" and striving for causal explanations and generalizability, especially in combination with quantitative data (e.g., Hammersley, 2008b). Qualitative approaches that manifest some or all of these characteristics include grounded theory (Glaser, 2000), qualitative comparative analysis (Ragin, 1987, 2000), and realism (Miles et al., 2014). Their common aim is to use qualitative methods to describe and explain phenomena as accurately and completely as possible so that their descriptions and explanations correspond as closely as possible to the way the world is and actually operates (Reynolds et al., 2011). Government agencies supporting qualitative research (e.g., the U.S. Government Accounting Office, the National Science Foundation, or the National Institutes of Health) usually operate within this traditional scientific framework.

SIDEBAR

THE ROOTS OF TRADITIONAL SOCIAL SCIENCE CRITERIA APPLIED TO QUALITATIVE INQUIRY

An emphasis on valid and reliable knowledge, as generated by neutral researchers utilizing the scientific method to discover universal Truth, reflects an epistemology commonly referred to as positivism. Historically, social scientists understood positivism as reflected in a "realist ontology, objective epistemology, and value-free axiology." Few, if any, qualitative researchers currently subscribe to an absolute faith in positivism, however. Many postpositivists, or researchers who believe that achievement of objectivity and value-free inquiry are not possible, nonetheless embrace the goal of production of generalizable knowledge through realist methods and minimization of researcher bias, with objectivity as a "regulatory ideal" rather than an attainable *goal*. In short, postpositivism does not embrace naive *belief* in pure scientific truth; rather, qualitative research conducted in a strict postpositivist tradition utilizes precise, prescribed processes and produces *social* scientific reports that enable researchers to make generalizable claims about the social phenomenon within particular populations under examination.

Postpositivists commonly utilize qualitative methods that bridge quantitative methods, in which researchers conduct an inductive analysis of textual data, form a typology grounded in the data (as contrasted with a preexisting, validated typology applied to new data), use the derived typology to sort data into categories, and then count the frequencies of each theme or category across data. Such research typically emphasizes validity of the coding schema, inter-coder reliability, and careful delineation of procedures, including random or otherwise *sys*tematic *sampling* of texts. Content analyses of media typify this approach. (Ellingson, 2011, pp. 596, 598; within-quote references omitted)

2. Social Construction and Constructivist Criteria

> **What is perceived as real is real in its consequences.**
>
> —The Thomas theorem

Social construction, constructivist, and "interpretivist" perspectives have generated new language and concepts to distinguish quality in qualitative research (e.g., Glesne, 1999, pp. 5–6). Lincoln and Guba (1986) proposed that constructivist inquiry demanded different criteria from those inherited from traditional social science. They suggested "credibility as an analog to internal validity, transferability as an analog to external validity, dependability as an analog to reliability, and confirmability as an analog to objectivity." In combination, they viewed these criteria as addressing "trustworthiness (itself a parallel to the term *rigor*)" (pp. 76–77). They went on to emphasize that naturalistic inquiry should be judged by dependability (a systematic process systematically followed) and authenticity (reflexive consciousness about one's own perspective, appreciation for the perspectives of others, and fairness in depicting constructions in the values that undergird them). They viewed the social world (as opposed to the physical world) as socially, politically, and psychologically constructed, as are human understandings and explanations of the physical world. They advocated triangulation to capture and report multiple perspectives rather than seek a singular truth. The team of researchers who reviewed approaches to assessing quality in qualitative research found that "the post-positivist criteria developed by Lincoln and Guba, based around the construct of 'trustworthiness,' were referenced frequently and appeared to be the basis upon which a number of authors made their recommendations for improving quality of qualitative research" (Reynolds et al., 2011).

Constructivists embrace subjectivity as a pathway deeper into understanding the human dimensions of the world in general as well as whatever specific phenomena they are examining (Peshkin, 1985, 1988, 2000a,b). They're more interested in deeply understanding specific cases within a particular context than in hypothesizing about generalizations and causes across time and space. Indeed, they are suspicious of causal explanations and empirical generalizations applied to complex human interactions and cultural systems. They offer perspective and encourage dialogue among perspectives rather than aiming at singular truths and linear predictions. Social constructivists' case studies, findings, and reports are explicitly informed by attention to praxis and reflexivity—that is, understanding how one's own experiences and background affect what one understands and how one acts in the world, including acts of inquiry. For an in-depth discussion of this perspective and its implications, see the *Handbook of Constructionist Research* (Holstein & Gubrium, 2008). Also see Chapter 3 pp. 121–126 for a much lengthier discussion of constructionism and constructivism.

Here are three examples of social construction as a framework for program evaluation.

1. The evaluation of a community development project in an ethnically and racially diverse neighborhood collected and reported stories from residents purposefully sampled to present a range of experiences and perspectives. The evaluation did not render judgments but was called a "multivocal evaluation" in which the diverse stories were used for dialogue and to enhance mutual understanding.
2. The evaluation of the international Paris Declaration on Development Aid included case studies that revealed the different perspectives and contexts within which aid is given and received. Donors (wealthier countries) and beneficiaries (poorer countries) experience different "realities." One purpose of the evaluation was to capture those different realities, including diverse experiences with the Paris Declaration principles, to facilitate dialogue on future international development policies and practices.
3. Social constructivism was the foundation of Nora Murphy's (2014) study of homeless youth. The 14 case studies showed diverse experiences and perspectives on homelessness. The study concluded that the unique situation of each homeless youth meant that program responses needed to be socially constructed together with the youth to be meaningful to them and to build trusting adult–youth relationships. (For more on this evaluation, see pp. 194, 626–628.)

SIDEBAR

CONSTRUCTIVIST TRUSTWORTHINESS

The credibility of your findings and interpretations depends on your careful attention to establishing trustworthiness. Lincoln and Guba (1985) describe prolonged engagement (spending sufficient time at your research site) and persistent observation (focusing in detail on those elements that are most relevant to your study) as critical in attending to credibility. "If prolonged engagement provides scope, persistent observation provides depth" (p. 304). With each, time is a major factor in the acquisition of trustworthy data. Time at your research site, time spent interviewing, and time building sound relationships with respondents all contribute to trustworthy data. When a large amount of time is spent with your research participants, they less readily feign behavior or feel the need to do so; moreover, they are more likely to be frank and comprehensive about what they tell you. Lincoln and Guba posited four constructivist criteria as parallel to but distinct from traditional research criteria:

> First, *credibility* (parallel to internal validity) addressed the issue of the inquirer providing assurances of the fit between respondents' views of their life ways and the inquirer's reconstruction and representation of same. Second, *transferability* (parallel to external validity) dealt with the issue of generalization in terms of case-to-case transfer. It concerned the inquirer's responsibility for providing readers with sufficient information on the case studied such that readers could establish the degree of similarity between the case studied and the case to which findings might be transferred. Third, dependability (parallel to reliability) focused on the process of the inquiry and the inquirer's responsibility for ensuring that the process was logical, traceable, and documented. Fourth, *confirmability* (parallel to objectivity) was concerned with establishing the fact that the data and interpretations of an inquiry were not merely figments of the inquirer's imagination. It called for linking assertions, findings, interpretations, and so on to the data themselves in readily discernible ways. For each of these criteria, Lincoln and Guba also specified a set of procedures that could be used to meet the criteria. For example, auditing was highlighted as a procedure useful for establishing both dependability and confirmability, and member check and peer debriefing, among other procedures, were defined as most appropriate for credibility.
>
> In *Fourth Generation Evaluation* (1989), Guba and Lincoln reevaluated this initial set of criteria. They explained that trustworthiness criteria were parallel, quasi-foundational, and clearly intended to be analogs to conventional criteria. Furthermore, they held that trustworthiness criteria were principally methodological criteria and thereby largely ignored aspects of the inquiry concerned with the quality of outcome, product, and negotiation. Hence, they advanced a second set of criteria called *authenticity criteria*, arguing that this second set was better aligned with the constructivist epistemology that informed their definition of qualitative inquiry. (Schwandt, 2007, pp. 299–300)

Continual alertness to your own biases and subjectivity (reflexivity) also assists in producing more trustworthy interpretations. Consider your subjectivity within the context of the trustworthiness of your findings. Ask yourself a series of questions: Whom do I not see? Whom have I seen less often? Where do I not go? Where have I gone less often? With whom do I have special relationships, and in what light would they interpret phenomena? What data-collecting means have I not used that could provide additional insight? Triangulated findings contribute to credibility. Triangulation may involve the use of multiple data collection methods, sources, investigators, or theoretical perspectives. To improve trustworthiness, you can also consciously search for negative cases.

Alternative Criteria Review

Exhibit 9.7 (pp. 680–681) presented seven different sets of criteria for judging the quality of qualitative studies. This module reviewed the first two sets of criteria: (1) traditional scientific research criteria versus (2) constructivist criteria. Constructivist criteria emerged from the critique that traditional scientific research criteria were based on quantitative and experimental design thinking that, by the very nature of using those criteria for defining quality, led to qualitative studies being judged inferior. The next module makes the issue of judging quality even more complicated by adding five more sets of alternative and competing criteria.

A Realist Views a Constructivist Proposal

Alternative and Competing Criteria, Part 2

Artistic, Participatory, Critical Change, Systems, Pragmatic, and Mixed Criteria

> **The moral and social yearnings of fully realized human beings are not reducible to universal laws and cannot be studied like physics.**
>
> —Brooks (2010, p. A27)

This module continues the presentation and discussion of seven alternative sets of criteria for judging the quality of qualitative studies. The previous module covered (1) traditional social science research criteria and (2) social construction and constructivist criteria. This module covers (3) artistic and evocative criteria, (4) participatory and collaborative criteria, (5) critical change criteria, (6) systems and complexity criteria, and (7) pragmatic and utilization-focused criteria. We'll then examine mixing criteria.

3. Artistic and Evocative Criteria

> **TRUTH is visceral, palpable, sensuous, wrenching, hormonal, cognitive, cathartic, lyrical, contextual, awakening, fleeting, universal, and debatable. In other words, truth is art.**
>
> —From Halcolm's *Ruminations*

Researchers and audiences operating from the perspective of traditional scientific research criteria emphasize the scientific nature of qualitative inquiry. Researchers and audiences who view the world through the lens of social construction emphasize qualitative inquiry as a particularly human form of understanding centered on the capacity of people and groups to construct meaning. That brings us to this third alternative, which emphasizes that human beings both think and feel. Traditional social science and constructivist inquiries focus on cognitive, logical, sense-making analyses. The artistic and evocative approaches to qualitative inquiry want to bring forth our emotional selves and do so by integrating art and science. Science makes us think. Great art makes us feel. From the perspective of artistic and evocation qualitative inquirers, great qualitative studies should evoke both understandings (cognition) and feelings (emotions).

> Persons are moved by emotion. . . . People are their emotions. To understand who a person is, it is necessary to understand emotion. . . . Emotions cut to the core of people. Within and through emotion people come to define the surface and essential, or core, meanings of who they are. Emotions and moods are ways of disclosing the world for the person. (Denzin, 2009, pp. 1–2)

Artistic and evocative criteria focus on aesthetics, creativity, interpretive vitality, and expressive voice. Case studies become literary works. Poetry or performance art may be used to enhance the audience's direct experience of the essence that emerges from analysis. Artistically oriented qualitative analysts seek to engage those receiving the work, to connect with them, move them, provoke them, and stimulate them. Creative nonfiction and fictional forms of representation blur the boundaries between what is "real" and what has been created to represent the essence of a reality. A literal presentation of reality, real (scientific) or perceived (constructivism), yields to artistically created reality. The results may be called creative syntheses, ideal-typical case constructions, scientific poetics, or any number of phrases that suggest the artistic emphasis. Artistic expressions of qualitative analysis strive to provide an experience with the findings where "truth" or "reality" is understood to have *a feeling dimension* that is every bit as important as the cognitive dimension. Such qualitative inquiry is explicitly *sensuous* (Stoller, 2004) and emotional (Denzin, 2009).

The performance art of *The Vagina Monologues* (Ensler, 2001), based on interviews with women about their experiences of coming of age sexually, but presented as theater, offers a prominent example. The audience feels as much as knows the truth of the presentation because of the essence it reveals. In the artistic tradition, the analyst's interpretive and expressive voice, experience, and perspective may become as central to

the work as depictions of others or the phenomenon of interest. Here are some examples of artistic and evocative approaches used in program evaluation.

- A program for low-income, pregnant, drug-addicted teenagers asked the young women to draw pictures of their hearts and tell what the pictures meant. The initial hearts were portrayed as wounded, knifed, torn, mangled, and tortured. Over a period of four months (four drawings and accompanying stories), the pictures showed some sunshine, flowers, rainbows, and, most striking, connections to other hearts. No perfect valentines. Not even close. But to look at those raw drawings was to see hearts healing.
- Theater for development is being used in Nigeria to engage community members in discussing preliminary results from an evaluation, with actors replaying scenarios relating to a program uncovered through fieldwork. By involving community members in role-play, the early findings can be verified or corrected based on the participants' experiences of the program, and new perspectives can be recorded, which might otherwise have remained dormant (Folorunsho, 2014).
- A team of educational evaluators concerted interviews with students and teachers about an international cross-cultural summer experience into a play dramatizing critical events and key learnings. The play was performed for the school board as the project's evaluation report.
- Photographs taken before and after Vietnam implemented a motorbike helmet law showed dramatic changes in compliance. (See Exhibit 8.20, pp. 609 ; for other examples of using visuals created to present evaluation findings, see Exhibits 8.21 through 8.27, pp. 610–619.)

Exhibit 9.8 shows how an interview transcript was converted into a poem when presenting the findings, all the better to give the reader a feel for what was said and the affect it carried.

EXHIBIT 9.8 From Interview Transcript to Poem: An Artistic and Evocative Presentation

In May 1994, Corrine Glesne (1997) interviewed Dona Juana, an 86-year-old professor in the College of Education at the University of Puerto Rico.

> That she chose a bird to represent her was no surprise. Standing 5 feet tall, very thin ("a problem all my life"), and with bright dark eyes, she was birdlike in appearance. Her office was a nest of books, papers, and folders in organized piles on her large desk, on the beige metal filing cabinets next to the door opposite her desk, in the wooden cabinet along the wall to the right of her desk, on the shelves below the window to her left, and on the two chairs before her desk. There was no sense of disorder, but rather an impression of an archive that would illuminate Dona Juana's 50 years in research and higher education. (p. 203)

Below is the poem Glesne (1997) created from the interview transcript, followed by a table showing the conversion from transcript to poem.

The Poem

That Rare Feeling

I am a flying bird
moving fast
seeing quickly
looking with the eyes of God,
from the tops of trees.

How hard for country people
picking green worms
from fields of tobacco,
sending their children to school,
not wanting them to suffer
as they suffer.
In the urban zone,
students worked at night
and so they slept in school.
Teaching was the real university.

So I came to study
to find out how I could help.
I am busy here at the university,
there is so much to do.
But the University
is not the Island.

I am a flying bird
moving fast, seeing quickly
so I can give strength,
so I can have that rare feeling
of being useful. (pp. 202–203)

Composing the Poem From the Interview Transcript

TRANSCRIPT		POETIC NARRATIVE CREATED
C (Corrine):	If I asked you to use a metaphor to describe yourself as a professor, what would you say you were like? Someone I asked said that she was a bridge and then she told me why. What metaphor comes to mind for you?	*Version 1: Chronologically and linguistically faithful to the transcript* I would be a flying bird. I want to move so fast so I can see quickly, everything. I wish I could look at the world with the eyes of God, to give strength to those that need . . . *Version 2: Draws from other sections of the interviews, takes more license with words.* I am a flying bird moving fast, seeing quickly, looking with the eyes of God from the tops of trees: How hard for country people picking green worms from fields of tobacco, sending their children to school, not wanting them to suffer as they suffer In the urban zone, students worked at night and so they slept in school. Teaching was the real university. So I came to study to find out how I could help. I am busy here at the university, there is so much to do. But the university is not the Island. I am a flying bird moving fast, seeing quickly so I can give strength, so I could have that rare feeling of being useful. (Glesne, 1997, p. 207)
J (Juana):	I would be a flying bird.	
C:	A flying bird. Tell me about it.	
	How are you a flying bird?	
J:	Because I want to move so fast.	
C:	Mrn-hmmm. Cover a lot of territory.	
J:	Yes. Yes.	
C:	Are you any kind of bird or just any bird?	
J:	Well, any bird because I don't want to mention some birds, some birds here are destructive.	
C:	Are what?	
J:	Are destructive. They destroy and I don't want to . . .	
C:	No, you don't want to be one of them. No. You're just a bird that moves fast.	
J:	That moves fast and sees from the tops of trees. So I can see quickly.	
C:	See quickly, see everything.	
J:	Everything.	
J:	So you can see me?	
C:	I can. I can see you, a flying bird.	
J:	I wish I could look at the world with the eyes of God.	
C:	With the eyes of what?	
J:	Of God, of that spiritual power that can give strength.	
C:	That can give strength? Strength?	
J:	Yes, to those that need.	

Crystallization

Sociologist Laurel Richardson (2000b) introduced *crystallization* as a criterion of quality in artistic and evocative qualitative inquiry, a replacement for triangulation as a criterion.

> The scholar draws freely on his or her productions from literary, artistic, and scientific genres, often breaking the boundaries of each of those as well. In these productions, the scholar might have different "takes" on the same topic, what I think of as a postmodernist deconstruction of triangulation. . . . In postmodernist mixed-genre texts, we do not triangulate, we *crystallize*. . . . I propose that the central image for "validity" for postmodern texts is not the triangle—a rigid, fixed, two-dimensional object. Rather, the central imaginary is the crystal, which combines symmetry and substance with

an infinite variety of shapes, substances, transmutations, multidimensionalities, and angles of approach. . . . Crystallization provides us with a deepened, complex, thoroughly partial, understanding of the topic. Paradoxically, we know more and doubt what we know. Ingeniously, we know there is always more to know. (p. 934)

Crystallization's roots can be traced to

the creative and courageous work of feminist methodologists who blasphemed the boundaries of art and science. . . .

Art and science do not oppose one another; they anchor ends of a continuum of methodology, and most of us situate ourselves somewhere in the vast middle ground. When scholars argue that we cannot include narratives alongside analysis or poems within grounded theory, they operate under the assumption that art and science negate one another and hence are incompatible, rather than merely differ in some dimensions. . . . My explanation of crystallization assumes a basic understanding of the complexities involved in combining methods and genres from across regions of the continuum. (Ellingson, 2009, pp. 3, 5)

4. Participatory and Collaborative Criteria

To be human is to engage in interpersonal dynamics. *Inter:* between. *Personal:* people. *Dynamics:* forces that produce activity and change. Combining these definitions, interpersonal dynamics are the forces between people that lead to activity and change. Whenever and wherever people interact, these dynamics are at work.

—King and Stevahn (2013, p. 2)
Interactive Evaluation Practice

Participatory and collaborative qualitative inquiries have four purposes and justifications:

1. ***Values premise:*** The right way to inquire into a phenomenon of interest is to do it with the people involved and affected. This means doing research and evaluation *with* as opposed *to* people. It means engaging them as fellow inquirers and coresearchers rather than as research subjects.

SIDEBAR

RIGOR IN ARTISTIC AND EVOCATIVE CRYSTALLIZATION

One of the most helpful (albeit not foolproof) ways to enhance your account and ward off editorial defensiveness toward creative analytic work, in general, and crystallization, in particular, is to be absolutely clear about what you did (and did not do) in producing your manuscript. This includes data collection, analysis, and especially choices made about representation. . . . By explaining my process, I help alleviate suspicions that I took an "anything goes" sloppy attitude toward constructing my representation.

While some colleagues may not like or approve of what you did no matter how you explain it, concise, explicit details of your process make it more difficult for them to dismiss it as careless or random. Accounting for your process (even in an appendix or endnote) constitutes an important nod toward methodological rigor. As many have posited, engaging in creative analytic work should be no less rigorous, exacting, and subject to strict standards of peer evaluation. . . . Moreover, such a roadmap assists others who may seek to follow your lead. . . .

Some suggestions on issues to own:

- Explain choices you made in composing narratives, poems, or other artistic work; in other words, how did you get from data to text?
- Describe your standpoint vis-à-vis your topic, not just what it is, but (at least some of) how it shapes your interactions with your data (e.g., I am a cancer survivor studying clinics so I tend to be more empathetic with patients than health care providers; I am a feminist so I pay a lot of attention to power dynamics).
- Indicate your awareness of and response to ethical considerations about voice, privacy, and responsibility to others. What steps did you take to ensure participant confidentiality? To privilege participants' voices? Consider how your work might be read in ways that do not reflect your intentions—for example, what quotes from participants could be taken out of context and used as justification for blaming the victim?—and surround vulnerable voices with preemptory statements that make it more difficult for oppositional forces to excerpt and reinterpret their meaning in regressive ways.
- Detail your analytic procedures. . . . Even if you construct a unique, outside-the-box artistic creation, you should explain your methodology and cite some sources to contextualize your work. Again, this need not interfere with your aesthetic goals; details should be concise and can be placed in an appendix, footnote, or even a separate piece altogether. The goal is to reveal crystallized projects as embodied, imperfect, insightful constructions rather than as immaculate end products. (Ellingson, 2009, pp. 199–120)

2. ***Quality premise:*** Data will be better when people who are the focus of the inquiry willingly participate, understand the nature of the inquiry, and agree with the importance of the study. Interviews will be richer and more detailed. Observations will be open and unguarded. Documents will be readily available. Data are better.
3. ***Reciprocity premise:*** Researchers get data, publications, knowledge, and career advancement from research and evaluation studies. Those who are the focus of inquiry should benefit as well. As coresearchers, through participation in the inquiry, they learn research skills, learn to think more systematically, and gain knowledge that they can use for their own purposes.
4. ***Utility premise:*** In program evaluation and action research inquiries, the findings are more likely to be useful—and actually used—when those who must act on the findings collaborate in generating and interpreting them.

From the classic articulation and justification of *Participatory Action Research* by William Foote Whyte (1989, 1991) to methods and facilitation guides on how to actually do it (Caister et al., 2011; Hacker, 2013; King & Stevahn, 2013; Pyrch, 2012; Taylor, Suarez-Balcazar, Forsyth, & Kielhofner, 2006), participatory and collaborative engagement has been a major approach to qualitative inquiry. When conducting research in a collaborative mode, professionals and nonprofessionals become coresearchers. Participatory action research encourages collaboration within a mutually acceptable inquiry framework to understand and/or solve organizational or community problems. Chapter 4 includes an in-depth discussion of participatory and collaborative approaches (pp. 213–222), including Exhibit 4.13, Principles of Fully Participatory and Genuinely Collaborative Inquiry (p. 222).

Here's an example of a participatory and collaborative qualitative inquiry. Robin Boylorn (2008) studied the experiences of black Southern women. She recruited a group of participants from the community in which she had grown up and invited them to share stories about their experiences and lives growing up and raising families in the rural South. She facilitated their interactions together as co-investigators so that they felt "equally invested and equally involved in the process of collecting, writing, interpreting, and editing the stories they wrote" (p. 600). She shared her experiences with the participants, and together they compared and contrasted their ideas and experiences.

> Their involvement began during the early stages of recommending other participants and retelling stories in individual and group settings to ensure adequate information was available. As co-investigators their stories were instrumental in establishing and representing a corporate set of themes and experiences. Though the co-researchers in this project were not involved in the writing stages, they did have the opportunity to respond to the stories the author wrote, offering their unique perspectives and feedback as participants in the research and characters in the stories. The resulting research project is a collaboration between the researcher and the researched, including participants as co-researchers. (p. 600)

SIDEBAR

INTERPERSONAL VALIDITY

Educational evaluator Karen Kirkhart (1995) coined the term *interpersonal validity*: the extent to which an evaluator is able to relate meaningfully and effectively to individuals in the evaluation setting. The interpersonal factor that undergirds interpersonal validity highlights the competence of a participatory evaluator or researcher do two things: (1) interact with people constructively throughout the framing and implementation of an inquiry and (2) create activities and conditions conducive to positive interactions among participants. The *interpersonal factor* is concerned with creating, managing, and ultimately mastering the interpersonal dynamics that make a collaborative inquiry possible and inform its findings. One is concerned with eventual use, the other with establishing buy-in among participants and a valid inquiry process. (King & Stevahn, 2013, p. 6)

5. Critical Change Criteria

We are distressed by underprivilege. We see gaps among privileged patrons and managers and staff and underprivileged participants and communities. . . . We are advocates of a democratic society.

—Robert Stake (2004, pp. 103–107)
Qualitative evaluation pioneer

How Far Dare an Evaluator Go Toward Saving the World?

Those engaged in qualitative inquiry as a form of critical analysis aimed at social and political change

eschew any pretense of open-mindedness or objectivity; they take an activist stance. Critical change inquiry aims to critique existing conditions and through that critique bring about change. Critical change criterion is derived from *critical theory*, which frames and engages in qualitative inquiry with an explicit agenda of elucidating power, economic, and social inequalities. The "critical" nature of critical theory flows from a commitment to go beyond just studying society for the sake of increased understanding. Critical change researchers set out to use inquiry to critique society, raise consciousness, and change the balance of power in favor of those less powerful. Influenced by Marxism, informed by the presumption of the centrality of class conflict in understanding community and societal structures, and updated in the radical struggles of the 1960s, *critical theory* provides both philosophy and methods for approaching research and evaluation as fundamental and explicit manifestations of political praxis (connecting theory and action), and as change-oriented forms of engagement.

> Critical social science and critical social theory attempt to understand, analyze, criticize, and alter social, economic, cultural, technological, and psychological structures and phenomena that have features of oppression, domination, exploitation, injustice, and misery. They do so with a view to changing or eliminating these structures and phenomena and expanding the scope of freedom, justice, and happiness. The assumption is that this knowledge will be used in processes of social change by people to whom understanding their situation is crucial in changing it. (Bentz & Shapiro, 1998, p. 146; Kincheloe & McLaren, 2000)

Critical change has three interconnected elements: (1) inquiry into situations of social injustice, (2) interpretation of the findings as a critique of the existing situation, and (3) using the findings and critique to mobilize and inform change.

> Critical theory looks at, exposes, and questions hegemony—traditional power assumptions held about relationships, groups, communities, societies, and organizations—to promote social change. Combined with action research, critical theory questions the assumed power that researchers typically hold over the people they typically research. Thus, critical action research is based on the assumption that society is essentially discriminatory but is capable of becoming less so through purposeful human action.
>
> Critical action research also assumes that the dominant forms of professional research are discriminatory and must be challenged. Critical action research takes the concept of knowledge as power and equalizes the generation of, access to, and use of that knowledge. Critical action research is an ethical choice that gives voice to, and shares power with, previously marginalized and muted people. (Davis, 2008, p. 140)

Critical change criteria apply to a number of specialized areas of qualitative inquiry (Given, 2008, pp. 139–179; Schwandt, 2007, pp. 50–55):

Critical ethnography	Critical discourse analysis	Critical realism
Critical education studies	Critical hermeneutics	Critical research
Critical arts-based inquiry	Critical humanism	Critical theory
Critical race theory	Critical pragmatism	Critical action research
Critical pedagogy	Critical social science	Critical systems analysis

In addition, *feminist inquiry* often includes an explicit agenda of bringing about social change (e.g., Benmayor, 1991; Brisolara, Seigart, & SenGupta, 2104; Hesse-Biber, 2013; Podems, 2014b). *Liberation research* and *empowerment evaluation* derive, in part, from Paulo Freire's philosophy of praxis and liberation education, articulated in his classics *Pedagogy of the Oppressed* (1970) and *Education for Critical Consciousness* (1973), still sources of influence and debate (e.g., Glass, 2001). Barone (2000) aspires to "emancipatory educational storysharing" (p. 247). Qualitative studies informed by critical change criteria range from largely intellectual and research-oriented approaches that aim to expose injustices to more activist forms of inquiry that actually engage in bringing about social change. Stephen Brookfield (2004) uses critical theory to illuminate adult education issues, trends, and inequities. Plummer (2011) integrates critical their and queer theory. Caruthers and Friend (2014) bring critical inquiry to online learning and engagement. Crave, Zaleski, and Trent (2014) emphasize the role of critical change in building a more equitable future through participatory program evaluation.

Here are two examples of critical change studies that would expect to be evaluated for quality by critical change criteria (Davis, 2008, p. 141):

1. Martin Diskin worked with policymakers and development agencies in Latin American studies to conduct what they called "power structure research," in which they exposed injustice as a strategy for building coalitions and motivating movements.

2. Christine Davis's ethnography of a children's mental health treatment team was an interdisciplinary research project involving the fields of communication studies, social work, and mental health. Conducted in partnership with community agencies, this research examined issues of power, marginalization, and control within these teams. It suggested a stance toward children and families that rejects the traditional hierarchical medical model of care and instead treats them as unique valuable humans and as equal partners in treatment.

Consequential validity as a critical change criterion for judging a research design or instrument makes the social consequences of its use a value basis for assessing its credibility and utility. Thus, standardized achievement tests are criticized because of the discriminatory consequences for minority groups of educational decisions made with "culturally biased" tests. Consequential validity asks for assessments of who benefits and who is harmed by an inquiry, measurement, or method (Brandon, Lindberg, & Wang, 1993; Messick, 1989; Shepard, 1993).

SIDEBAR

A QUALITATIVE MANIFESTO: A CALL TO ARMS

—Norman K. Denzin (2010)

> The social sciences . . . should be used to improve quality of life. . . . For the oppressed, marginalized, stigmatized and ignored . . . and to bring about healing, reconciliation and restoration between the researcher and the researched.
>
> —Stanfield (2006, p. 725)

Mills wanted his sociology to make a difference in the lives that people lead. He challenged persons to take history into their own hands. He wanted to bend the structures of capitalism to the ideologies of radical democracy. . . .

I want a critical methodology that enacts its own version of the sociological imagination. Like Mills, my version of the imagination is moral and methodological. And like Mills, I want a discourse that troubles the world, understanding that all inquiry is moral and political.

This book is an invitation and a call to arms. It is directed to all scholars who believe in the connection between critical inquiry and social justice (Denzin, 2010, p. 10).

Qualitative inquiry can contribute to social justice in the following ways:

1. It can help identify different definitions of a problem and/or a situation that is being evaluated with some agreement that change is required. It can show, for example, how battered wives interpret the shelters, hotlines, and public services that are made available to them by social welfare agencies. Through the use of personal experience narratives, the perspectives of women and workers can be compared and contrasted.
2. The assumptions, often belied by the facts of experience, that are held by various interested parties—policy makers, clients, welfare workers, online professionals—can be located and shown to be correct, or incorrect (Becker, 1967, p. 239).
3. Strategic points of intervention into social situations can be identified. Thus, the services of an agency and a program can be improved and evaluated.
4. It is possible to suggest "alternative moral points of view from which the problem, the policy and the program can be interpreted and assessed" (see Becker, 1967, pp. 239–240). Because of its emphasis on experience and its meanings, the interpretive method suggests that programs must always be judged by and from the point of view of the persons most directly affected.
5. The limits of statistics and statistical evaluations can be exposed with the more qualitative, interpretive materials furnished by this approach. Its emphasis on the uniqueness of each life holds up the individual case as the measure of the effectiveness of all applied programs. (Denzin, 2010, pp. 24–25)

6. Systems Thinking and Complexity Criteria

In a finite game, it is easy to make sense. Everyone agrees on the goal; the rules are known; and the field of play has clear boundaries. Baseball, football, and bridge are examples of finite games. At one time in the not-so-distant past we expected careers, marriages, parenthood, education, and citizenship to be finite games. When everyone agrees on the rules, and the consequences of our actions are undeniable, responsible people plan for what they want, take steps to achieve it, and enjoy the fruits of their labor. We know what it takes to make sense in a finite game.

> Most of us realize we are playing in a very different game. We are playing in an infinite game. In which the boundaries are unclear or nonexistent, the scorecard is hidden, and the goal is not to win but to keep the game in play. There are still rules, but the rules can change without notice. There are still plans and playbooks, but many games are going on at the same time, and the winning plans can seem contradictory. There are still partners and opponents, but it is hard to know who is who, and besides that, the "who is who" changes unexpectedly.
>
> —Glenda Eoyang and Royce Holladay (2013, p. 4)

Adaptive Action: Leveraging Uncertainty in Your Organization

Studying "infinite games" in highly dynamic situations characterized by uncertainty and rapid change creates special challenges for qualitative inquiry. Systems thinking and complexity concepts offer a framework for studying such situations, tools for both inquiry and "coping with chaos" (Eoyang, 1997), and criteria for deciding whether such studies are of high quality. To be credible to systems thinkers and complexity scientists, the qualitative inquiry must capture, describe, map, and analyze, and map systems of interests; must attend to interrelationships, capture diverse perspectives, attend to emergence, and be sensitive to and explicit about boundary implications; and must document nonlinearities, adapt the inquiry in the face of uncertainties, and describe systems changes and their implications. In so doing, the explanatory approach moves from attribution to contribution analysis (see pp. 596–597 in Chapter 8).

Chapter 3 discussed systems theory and complexity theory as distinct, though intersecting, theoretical frameworks (see pp. 139–151). Exhibit 3.14 presents complexity theory concepts and qualitative inquiry implications (pp. 147–148). Exhibit 3.16 presents the relationship of systems theory to complexity theory (p. 150).

- ***Systems theory inquiry questions:*** How and why does this system function as it does? What are the system's boundaries and interrelationships, and how do these affect perspectives about how and why the system functions as it does?
- ***Complexity theory inquiry question:*** How can the emergent and nonlinear dynamics of complex adaptive systems be captured, illuminated, and understood?

For my purpose here, namely, differentiating distinct sets of criteria by which to judge the quality and credibility of various approaches to qualitative inquiry, the core systems and complexity dimensions can be integrated, as they are in Exhibit 9.7. That said, the systems field exemplifies the challenge of settling on some definitive set of quality criteria for judging qualitative inquiry, especially using a systems and complexity framing, because there are multiple approaches within the systems field (e.g., Hieronymi, 2013), each of which would assert and favor particular criteria unique to that perspective.

> Systemic inquiry covers a wide range methodologies, methods, and techniques with a strong focus on the behaviors of complex situations and the meanings we draw from those situations. It spans both the qualitative and quantitative research method domains but also includes approaches that fit neither category nor both categories. . . .
>
> Any attempt to summarize a trans discipline like systemic inquiry is fraught with difficulties. Despite relatively simple origins, the field has sprawled many directions so that no single, universally accepted theory has emerged, and neither are there universally agreed definitions of basic concepts such as what is and what is not a system. Although we will find many definitions in the systems literature, many authors argue that single fixed definitions promote the kind of reductionist thinking that runs counter to systemic principles. Instead, they argue, the field should promote debates around methodological principles to create learning rather than fixed definitions—what Kurt Richardson calls "critical pluralism." (Williams, 2008, p. 858)

Thinking Systemically

So as not to get lost in or overwhelmed by different approaches to systems, let me close this section with an example grounded in the basics of attending to interrelationships, boundaries, perspectives, and emergence. An exemplar of applying systems thinking to understand an issue is the analysis done by Christopher Wells (2012) of the role and impact of the automobile in the United States. His analysis begins *before* there were automobiles (what complexity theorists call *initial conditions*). He examines emergent land use patterns in the nineteenth century, sanitation problems in cities, the development of agricultural markets, the role of horses in transportation, the influence of train routes, the challenges of riding bicycles on rutted and muddy roads, the function of farmers in maintaining roads along their farms, population growth, and many other factors that established the initial conditions that automobiles emerged into. To understand the automobile in American society, culture, politics, and economics, you must look at the systems before the automobile existed (transportation, commerce, public health, political

jurisdictions, land use, and community values as starting places) and continue to examine those systems and their interactions through to the present day. The irony is that engaging and thinking through those complex interactions yields extraordinary clarity.

7. Pragmatic, Utilization-Focused Criteria

> *Usefulness!* It is not a fascinating word, and the quality is not one of which the aspiring spirit can dream o' nights, yet on the stage it is the first thing to aim at.
>
> —Dame Ellen Terry (1847–1928)
> Leading Shakespearean actress in Britain
>
> How use doth breed a habit in a man.
>
> —William Shakespeare

It is intriguing to find a great Shakespearean actress lauding *usefulness* as a matter of prime concern in her performances. Based on her musings about what she aspired to, usefulness concerned using anything and everything at her disposal to bring the play to life and connect with the audience. This is, perhaps, an artistic and evocative view of usefulness, but it also connotes a practical twist that makes for a provocative introduction to our final set of quality criteria: *pragmatic, utilization-focused criteria.*

- Observations of a high school cafeteria revealed substantial food waste. Interviews showed why the students were so dissatisfied with the food offered. The school had recently experienced an influx of immigrants from Asian countries, where people preferred rice rather than potatoes and bread. The results were used by school officials and the student council to advocate for more culturally appropriate food. Their efforts were successful.
- An early-childhood parent education program was experiencing a high dropout rate. Fewer than half the parents who started the program completed it. Interviews with the dropouts revealed that the program materials being used were academic and difficult for poorly educated parents to understand. Materials were revised to be more accessible and appropriate for parents with lower reading skills.
- The agricultural extension service serving a remote rural area in West Africa had very poor attendance at field trips aimed at helping farmers improve their basic growing practices for the subsistence crops sorghum and millet. Interviews with farmers revealed that they received no advance notice when the field training would occur. A radio news program agreed to announce extension field visits. Attendance increased significantly.

These are examples of simple inquiries aimed at providing practical and useful information to solve immediate problems. The pragmatic, utilization-focused criteria emphasize qualitative data generated to solve problems and inform decisions. This means focusing the inquiry on informing action and decisions. To be useful, specific intended users must be identified and their information needs met. Interactive engagement with intended users enhances relevance and use. Findings and feedback are timed to support use. Findings must be actionable and results understandable. The methods used need to be credible to those who will use the findings. Epistemologically, the orientation of pragmatic qualitative inquiry is that what is useful is true.

Pragmatic, utilization-focused inquiry begins with the premise that studies should be judged by their

SIDEBAR

DIVERSE METHODS BASED ON SYSTEMS AND QUALITY CONCEPTS

Systems and complexity concepts manifest nuances of difference under varying application frameworks:

- ***System dynamics:*** Focuses on the interrelationships between components of a situation, especially the consequences of feedback and delay
- ***Viable systems:*** Explores relationships that support an organization's viability within its environment
- ***Soft systems methodology:*** Looks at a situation from multiple viewpoints to understand and anticipate both interactions and unanticipated consequences
- ***Critical systems heuristics:*** Focuses on ethical issues, marginalization of people, and ideas of power and coercion
- ***Activity systems:*** Draws on cultural-historical activity theory to identify and track roles, tools, past features and dynamics, contradictions, tensions, conflicts, disturbances, innovations, processes, and learning opportunities
- ***Complex adaptive systems:*** Independent and interdependent elements or agents adapting to each other, self-organizing and emergent patterns, and nonlinear dynamics
- ***Network analysis:*** Examines dynamic interactions, connectivity, processes, and outcomes among a group or system of interconnected people or things. (Network Impact and Center for Evaluation Innovation, 2014)

SOURCES: Williams (2005) and Williams and Hummelbrunner (2011).

utility and actual use; therefore, evaluators and researchers should facilitate the inquiry process and design any study with careful consideration of how everything that is done, *from beginning to end*, will affect use. Use concerns how real people in the real world apply findings and experience the inquiry process. Therefore, the *focus* is on *intended use by intended users*. Since no study can be value-free, utilization-focused inquiry answers the question of whose values will frame the study by working with clearly identified, primary intended users who have the responsibility to apply findings and take action (Patton, 2008, 2012a).

SIDEBAR

PRAGMATIC EVALUATION STANDARDS

The evaluation profession has adopted standards that call for evaluations to be useful, practical, ethical, accurate, and accountable (Joint Committee on Standards, 2010). In the 1970s, as evaluation was just emerging as a field of professional practice, many evaluators took the position of traditional researchers that their responsibility was merely to design studies, collect data, and publish findings; what decision makers did with those findings was not their problem. This stance removed from the evaluator any responsibility for fostering use and placed all the "blame" for nonuse or underutilization on decision makers. Moreover, before the field of evaluation identified and adopted its own standards, criteria for judging evaluations could scarcely be differentiated from criteria for judging research in the traditional social and behavioral sciences, namely, technical quality and methodological rigor. Utility was largely ignored. Methods decisions dominated the evaluation design process. Validity, reliability, measurability, and generalizability were the dimensions that received the greatest attention in judging evaluation research proposals and reports. Indeed, evaluators concerned about increasing a study's usefulness often called for ever more methodologically rigorous evaluations to increase the validity of findings, thereby supposedly compelling decision makers to take findings seriously.

By the late 1970s, however, program staff and funders were becoming openly skeptical about spending scarce funds on evaluations that they couldn't understand and/or found irrelevant. Evaluators were being asked to be "accountable," just as program staff were supposed to be accountable. The questions emerged with uncomfortable directness: Who will evaluate the evaluators? How will evaluation be evaluated? It was in this context that professional evaluators began discussing standards.

The most comprehensive effort at developing standards was hammered out over five years by a 17-member committee appointed by 12 professional organizations with input from hundreds of practicing evaluation professionals. Just prior to publication, Dan Stufflebeam, chair of the committee, summarized the results as follows:

> The standards that will be published essentially call for evaluations that have four features. These are *utility, feasibility, propriety* and *accuracy*. And I think it is interesting that the Joint Committee decided on that particular order. Their rationale is that an evaluation should not be done at all if there is no prospect for its being useful to some audience. Second, it should not be done if it is not feasible to conduct it in political terms, or practicality terms, or cost effectiveness terms. Third, they do not think it should be done if we cannot demonstrate that it will be conducted fairly and ethically. Finally, if we can demonstrate that an evaluation will have utility, will be feasible and will be proper in its conduct, then they said we could turn to the difficult matters of the technical adequacy of the evaluation. (Stufflebeam, 1980, p. 90)

High-Stakes Debate: What Counts as Credible Evidence, and by What Criteria Shall Credibility Be Judged?

The seven frameworks just reviewed show the range of criteria that can be brought to bear in judging a qualitative study. They can also be viewed as "angles of vision" or "alternative lenses" for expanding the possibilities available, not only for critiquing inquiry but also for undertaking it. What is most important to understand is that researchers and evaluators attending to and operating with any one of the seven different sets of quality criteria will (a) ask different questions, (b) use different methods, (c) follow different analytical processes, (d) report their findings in different ways, and (e) aim their claims of credibility to different audiences. These are not just academic distinctions. The differences are far from trivial. Quite the contrary, the different orientations have far-reaching implications for every aspect of inquiry. These different quality criteria constitute the underpinnings of significantly different ways of engaging in qualitative inquiry. At the heart of all scientific debate throughout history has been this burning question: *What counts as credible evidence and by what criteria shall credibility be judged?*

Nor is the debate about what counts as credible evidence just a matter of contention among scientists. Policymakers and politicians have gotten involved. It is down-and-dirty politics with millions of dollars in government-funded and philanthropic-sponsored research at stake (Denzin & Giardina, 2006, 2008; Donaldson, Christie, & Mark, 2008; Scriven, 2008). This means that advocates of qualitative inquiry must understand and be prepared to enter the debate about the politics of evidence (e.g., Eyben, 2013; Nutley et al., 2013; Schorr, 2012). In so doing, understanding the variety of approaches to qualitative inquiry, and which approaches are legitimate by what criteria, will become part of the debate.

So make no mistake about it, advocates of one particular set of criteria are likely to be vociferous critics of alternative criteria. Using the research process as an intervention to correct injustices and foment change is anathema to those who advocate traditional scientific research criteria as the only acceptable standards for judging quality. Those traditional criteria insist on a clear line of demarcation between studying a phenomenon (basic research and independent, external evaluation) versus engaging in change through the research process (advocacy). On the other hand, attempts to make traditional scientific research criteria the only legitimate approach to government-funded research are criticized as narrow-minded, self-serving political advocacy that constitutes "a conservative challenge to qualitative inquiry" (Denzin & Giardina, 2006, p. x). Constructivists generated their criteria of quality as a direct reaction to what they considered the gross inadequacies and methodological distortions of traditional scientific research criteria, which are essentially derived from the experimental/quantitative paradigm (see pp. 87–95). Thus, they systematically set out to replace traditional research criteria like validity and reliability with trustworthiness and authenticity (Lincoln & Guba, 1985, 1986). Advocates of artistic and evocative approaches attack both traditional research and constructivism as emotionally void. Traditional researchers have been disinclined to use participatory and collaborative approaches, sometimes believing that involving nonresearchers in research inevitably leads to poorer quality; in other cases, it's a matter of lacking incentives, capacity, or interest. Pragmatic, utilization-focused inquiries are attacked for being theoretically useless, unscholarly, and so practical as to be worthless for generating explanations or generalizations. Many traditional researchers don't even consider action research worthy of the name "research."

Choosing a Framework Within Which to Work

Which criteria you choose to emphasize in your work will depend on the purpose of your inquiry, the values and perspectives of the audiences for your work, and your own philosophical and methodological orientation. Operating within any particular framework and using any specific set of criteria will invite criticism from those who judge your work from a different framework and with different criteria. (For examples of the vehemence of such criticisms between those using traditional social science criteria and those using artistic narrative criteria, see Bochner, 2001; English, 2000.) Understanding that criticisms (or praise) flow from criteria can help you anticipate how to position your inquiry and make explicit what criteria to apply to your own work as well as what criteria to offer others given the purpose and orientation of your work.

The profession of program evaluation is a microcosm of these larger divisions. Program evaluation is a diverse, multifaceted profession manifesting many different models and approaches (Christie & Alkin, 2013; Fitzpatrick, Sanders, & Worthen, 2010; Funnell & Rogers, 2011; Patton, 2008; Stufflebeam, Madeus, & Kellaghan, 2000). All seven alternative quality criteria are advocated by various evaluation theorists, methodologists, and practitioners.

Any particular evaluation study has tended to be dominated by one set of criteria, with a second set as possibly secondary. For example, a primarily constructivist approach might add some artistic techniques as supporting methods. An evaluation dominated by the traditional scientific research approach might have a section dedicated to dealing with pragmatic issues. Exhibit 9.9 shows how the seven frameworks can be found in the approaches of various evaluation theorists, methodologists, and practitioners.

Clouds and Cotton: Mixing and Changing Perspectives

While each set of criteria manifest a certain coherence, many researchers mix and match approaches. The work of Tom Barone (2000), for example, combines aesthetic, political (critical change), and constructivist elements. Denzin's *Performance Ethnography* (2003) uses artistic and evocative approaches to foment and contribute to "radical social change, to economic justice, to a culture of politics that extends critical race theory and the principles of a radical democracy to all aspects of society" (p. 3). A team of evaluators collaborated to integrate constructivism, participatory evaluation, critical change, and a utilization focus (evaluations for improvement):

EXHIBIT 9.9 Alternative Quality Criteria Applied to Program Evaluation

Program evaluation is a diverse, multifaceted profession manifesting many different models and approaches (Alkin & Christie, 2013; Fitzpatrick, Sanders, & Worthen, 2010; Funnell & Rogers, 2011; Patton, 2008; Stufflebeam, Madeus, & Kellaghan, 2000). All seven alternative quality criteria are advocated by various evaluation theorists, methodologists, and practitioners.

QUALITY CRITERIA	PROGRAM EVALUATION FOCUS	LEADING CLASSIC TEXTS AND RESOURCES
1. Traditional scientific research criteria	Apply research methods to attribute documented outcomes to the intervention and generalize the findings.	Chen and Rossi (1987), Rossi, Lipsey, and Freeman (2004), and Silverman, Ricci, and Gunter (1990)
2. Social construction and constructivist criteria	Capture and report multiple perspectives on participants' experiences and diverse program outcomes.	Greene (1998, 2000), Guba and Lincoln (1981, 1989), Lincoln and Guba (1985), and Schwandt and Burgon (2006)
3. Artistic and evocative criteria	Connoisseurship evaluation: Use artistic representations to evoke participants' program experiences and judge a program's merit and worth.	Barone (2001, 2008), Eisner (1985, 1991), Knowles and Cole (2008), and Mathison (2009)
4. Participatory and collaborative criteria	Involve program staff and participants in evaluation to enhance use and build capacity for future evaluations.	Cousins and Chouinard (2012), Cousins and Earl (1992, 1995), King, Cousins, and Whitmore (2007), Greene (2006), and King and Stevahn (2013)
5. Critical change criteria	Use evaluation to address social justice, empower participants, and bring about change; support genuine democracy; and reduce power imbalances.	Fetterman (2000), Fetterman and Wandersman (2005), Fetterman, Kaftarian, and Wandersman (1996), Greene (2006), House and Howe (2000), Kirkhart (1995), and Mertens (1998, 1999, 2013)
6. Systems and complexity criteria	Understand programs through systems analysis and complexity concepts; support program innovation and adaptation; and evaluate systems change.	Eoyang and Holladay (2013), Jolley (2014), Mowles (2014), Patton (2011), Sterman (2006), Walton (2014), Williams and Hummelbrunner (2011), and Williams and Iman (2006)
7. Pragmatic, utilization-focused criteria	Get actionable answers to practical questions to support program improvement, guide problem solving, and enhance decision making, and ensure the utility and actual use of findings.	Alkin, Daillak, and White (1979), Davidson (2012), Patton (2008, 2012a), Rogers and Williams (2006), and Weiss (1977)

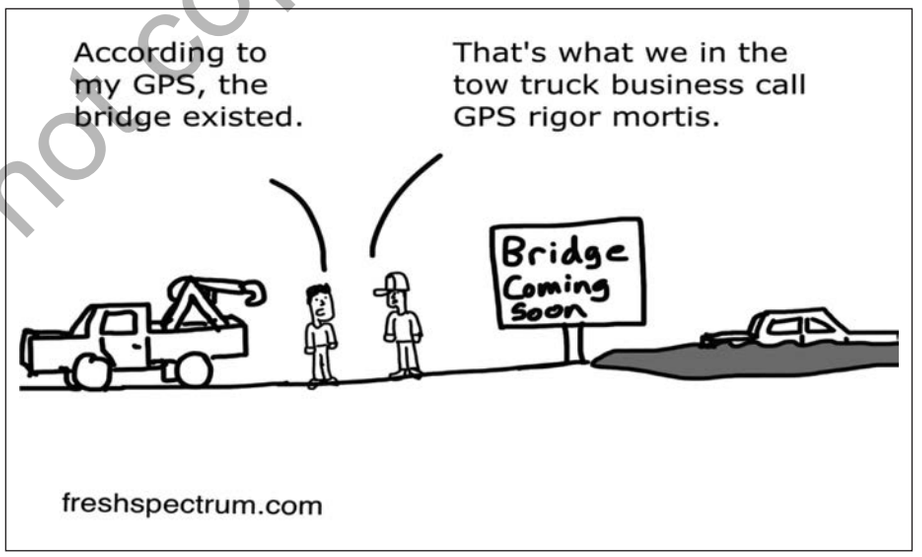
> *Evaluations for improvement, understanding lived experience, or advancing social justice are fundamentally participatory, involving key stakeholders in critical decisions about the evaluation's agenda, direction, and use.* Such a principle is rooted epistemologically in the importance of understanding multiple perspectives and experiences in evaluation, and also politically in the importance of democratic inclusion.
>
> —Whitmore et al. (2006, p. 341)

As an evaluator, I have worked with mixed criteria from all seven frameworks to match particular designs to the needs and interests of specific stakeholders and clients (Patton, 2008). But mixing and combining criteria means dealing with the tensions between them. After reviewing the tensions between traditional social science criteria and postmodern constructivist criteria, narrative researchers Lieblich, Tuval-Mashiach, and Zilber (1998) attempted "a middle course," but that middle course reveals the very tensions they were trying to supersede as they worked with one leg in each camp.

> We do not advocate total relativism that treats all narratives as texts of fiction. On the other hand, we do not take narratives at face value, as complete and accurate representations of reality. We believe that stories are usually constructed around a core of facts or life events, yet allow a wide periphery for freedom of individuality and creativity in selection, addition to, emphasis on, and interpretation of these "remembered facts." . . .
>
> Life stories are subjective, as is one's self or identity. They contain "narrative truth" which may be closely linked, loosely similar, or far removed from "historical truth." Consequently, our stand is that life stories, when properly used, may provide researchers with a key to discovering identity and understanding it—both in its "real" or "historical" core, and as narrative construction. (p. 8)

Traditional scientific research criteria and critical change criteria are polar opposites. The same study cannot aspire to independence, objectivity, and a primary focus on contributing to theory while also being deeply engaged in using the inquiry process to foment change and ameliorate oppression. Mixing methods (qualitative and quantitative) is one thing. Mixing criteria of quality is a bit more challenging, one might even say daunting. Certainly, constructivist, artistic, and participatory criteria can be intermingled. But traditional scientific research criteria are less amenable to comingling.

The remainder of this chapter will elaborate some of the most prominent of these competing criteria that affect judgments about the quality and credibility of qualitative inquiry and analysis. But it's not always easy to tell whether someone is operating from a realist, constructionist, artistic, activist, or evaluative framework. Indeed, the criteria can shift quickly. Consider this example. My six-year-old son, Brandon, was explaining a geography science project he had done for school. He had created an ecological display out of egg cartons, ribbons, cotton, bottle caps, and styrofoam beads. "These are three mountains and these are four valleys," he said, pointing to the egg cup arrangement. "And is that a cloud?" I asked, pointing to the big hunk of cotton. He looked at me, disgusted, as though I've just said about the dumbest thing he's ever heard. "That's a piece of cotton, Dad."

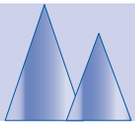


Foreshadowing research rigor mortis, MQP Rumination # 9 in the next module.

SOURCE: © Chris Lysy—freshspectrum.com

MODULE

80 Credibility of the Inquirer



The previous modules in this chapter have reviewed strategies for enhancing the quality and credibility of qualitative analysis: selecting appropriate criteria for judging quality, searching for rival explanations, explaining negative cases, triangulation, and keeping data in context. Technical rigor in analysis is a major factor in the credibility of qualitative findings. This section now takes up the issue of how the credibility of the inquirer affects the way findings are received.

One barrier to credible qualitative findings stems from the suspicion that the analyst has shaped findings according to her or his predispositions and biases. Whether this may have happened unconsciously, inadvertently, or intentionally (with malice and forethought) is not the issue. The issue is how to counter such a suspicion before it takes root. One strategy involves discussing your predispositions and making biases explicit, to the extent possible. This involves systematic and studious reflexivity (see pp. 70–74). Another approach is engaging in mental cleansing processes (e.g., *epoche* in phenomenological analysis, p. 575). Or one may simply acknowledge one's orientation as a feminist researcher (Podems, 2014b) or critical theorist and move on from there. The point is that you have to address the issue of your credibility.

The Researcher as the Instrument in Qualitative Inquiry

Because the researcher is the instrument in qualitative inquiry, a qualitative report should include some information about you, the researcher. What experience, training, and perspective do you bring to the study? Who funded the study and under what arrangements with you? How did you gain access to the study site and the people observed and interviewed? What prior knowledge did you bring to the research topic and study site? What personal connections do you have to the people, program, or topic studied? For example, suppose the observer of an Alcoholics Anonymous program is a recovering alcoholic. This can either enhance or reduce credibility depending on how it has enhanced or detracted from data gathering and analysis. Either way, the analyst needs to deal with it in reporting findings. In a similar vein, it is only honest to report that the evaluator of a family counseling program was going through a difficult divorce at the time of fieldwork.

No definitive list of questions exists that must be addressed to establish investigator credibility. *The principle is to report any personal and professional information that may have affected data collection, analysis, and interpretation*—either negatively or positively—in the minds of users of the findings. For example, health status should be reported if it affected one's stamina in the field. Were you sick part of the time? Let's say that the fieldwork for evaluation of an African health project was conducted over three weeks, during which time the evaluator had severe diarrhea. Did that affect the highly negative tone of the report? The evaluator said it didn't, but I'd want to have the issue out in the open to make my own judgment. Background characteristics of the researcher (e.g., gender, age, race, and/or ethnicity) may be relevant to report in that such characteristics can affect how the researcher was received in the setting under study and what sensitivities the inquirer brings to the issues under study.

In preparing to interview farm families in Minnesota, I began building up my tolerance for strong coffee a month before the fieldwork. Being ordinarily not a coffee drinker, I knew my body would be jolted by 10 to 12 cups of coffee a day doing interviews in farm kitchens. In the Caribbean, I had to increase my tolerance for rum because some farmer interviews took place in rum shops. These are matters of personal preparation—both mental and physical—that affect perceptions about the quality of the study. Preparation and training for fieldwork, discussed at the beginning of Chapter 6, should be reported as part of the study's methodology.

Reflexivity and Intellectual Rigor

(Othello to Iago, interpreting what it means for someone to mutter something while sleeping)

> **But this denoted a foregone conclusion.**
>
> —William Shakespeare
> (Othello to Iago, interpreting what it means for someone to mutter something while sleeping)

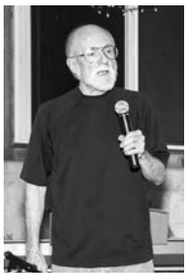
The credibility of qualitative inquiry is so closely connected to the credibility of the person or team conducting the inquiry that the quality of reflexivity and reflectivity offered in a report is a window into the thinking processes that are the bedrock of qualitative analysis. Essentially, reflexivity involves turning qualitative analysis on yourself. Who are you, and how has

how has who you are affected what you've found and reported in the study? This puts your intellectual rigor on display. The very notion of intellectual rigor connotes that as important as it is to employ systematic analytical strategies and techniques, the effectiveness and quality of those strategies and techniques depend on the quality of thinking that directs them. Which brings me to this chapter's rumination: Avoiding Research Rigor Mortis.

MQP Rumination # 9

Avoiding Research Rigor Mortis

I am offering one personal rumination per chapter. These are issues that have persistently engaged, sometimes annoyed, occasionally haunted, and often amused me over more than 40 years of research and evaluation practice. Here's where I state my case on the issue and make my peace.



Look for a pattern in what follows. See if you detect a theme.

Rigor (definition). Unyielding or inflexible; the quality of being extremely thorough, exhaustive, or accurate; being strict in conduct, judgment, and decision (*Oxford Dictionary*); scrupulous or inflexible accuracy or adherence (*Random House Dictionary*)

*Measurement rigor***.** The underlying psychometric properties of a measure and its ability to fully and meaningfully capture the relevant construct; the fact that data have been collected in essentially the same manner, across time, the program, and jurisdictions, adds methodological rigor; the reliability and validity of instruments (Weitzman & Silver, 2012)

*Research design rigor***.** The true experiment (randomized controlled trials) as the optimal (gold standard) design for developing evidence-based practice (Ross, Barkaoui, & Scott, 2007)

*Methodological rigor***.** Design elements that support strong causal attributions and analytical generalization (Chatterji, 2007; Coryn, Schröter, & Hanssen, 2009)

*Evaluation research rigor***.** Evidence testing the extent to which valid and reliable measures of program outcomes can be directly and confidently attributed to a standardized, high-fidelity, consistently implemented program intervention; the most rigorous evaluation is the randomized controlled trial (Chatterji, 2007; Henry, 2009; Ross et al., 2007; Rossi et al., 2004); "methodological rigor can be assessed from the evaluation plan and the quality of the evaluation's implementation" (Braverman, 2013, p. 101)

*Analytical rigor***.** "Meticulous adherence to standard process . . .; scrupulous adherence to established standards for the conduct of work" (Zelik, Patterson, & Woods, 2007, p. 1)

*Rigor mortis***.** Latin: *rigor* "stiffness," *mortis* "of death"—one of the recognizable signs of death, caused by chemical changes in the muscles after death, causing the limbs of the corpse to become stiff and difficult to move or manipulate

*Research rigor mortis***.** Rigid designs, rigidly implemented, then rigidly analyzed through standardized, rigidly prescribed operating procedures and judged hierarchically by standardized, rigid criteria, thereby manifesting *rigorism* at every stage

*Rigorism***.** Extreme strictness; no course may be followed that is contrary to doctrine (*Random House Dictionary*)

*Research rigorism***.** *Technicism*—reducing research to "the application of techniques or the following of rules" (Hammersley, 2008b, p. 31)

Did you find the pattern? Did you detect a theme? Read on for the *countertheme*. (A countertheme is like a counterfactual: a theme that might be dominant, even should be dominant, in an alternate universe where the dominant theme is not so *dominant*.)

The Problem

"The Problem of Rigor in Qualitative Research"—that's the title of a classic article (Sandelowski, 1986) and a common refrain in textbooks about research methods. The "problem," it turns out, is that by traditional and dominant definitions of rigor, qualitative methods are inferior. But different criteria for what constitutes methodological quality lead to different judgments about rigor, the central point of this chapter. "The 'problem of multiple standards' describes the inherent difficulties in selecting which, among many viable candidates, is *the* standard process to which performance should be compared" (Zelik, Patterson, & Woods, 2007, p. 2). Rigor begets credibility. Different criteria for what constitutes methodological quality and rigor will yield different judgments about credibility. That much is straightforward.

The larger problem, it seems to me, is the focus on methods and procedures as the basis for determining quality and rigor. The notion that methods are more or less rigorous decouples methods from context and the thinking process that determined what questions to ask, what methods to use,

(Continued)

(Continued)

what analytical procedures to follow, and what inferences to draw from the findings. Avoiding *research rigor mortis* requires rigorous thinking.

Rigorous Thinking

> No problem can withstand the assault of sustained thinking.
>
> —Voltaire (1694–1778)
> French philosopher

Rigorous thinking combines (a) critical thinking, (b) creative thinking, (c) evaluative thinking, (d) inferential thinking, and (e) practical thinking. *Critical thinking* demands questioning assumptions; acknowledging and dealing with preconceptions, predilections, and biases; diligently looking for negative and disconfirming cases that don't fit the dominant pattern; conscientiously examining rival explanations; relentlessly seeking diverse perspectives; and analyzing what and how you think, why you think that way, and the implications for your inquiry (Kahneman, 2011; Klein, 2011; Loseke, 2013).

Creative thinking invites putting the data together in new ways to see the interactions among separate findings more holistically; synthesizing diverse themes in a search for coherence and essence while simultaneously developing comfort with ambiguity and uncertainty in the messy, complex, and dynamic real work; distinguishing signal from noise while also learning from the noise; asking wicked questions that enter into the intersections and tensions between the search for coherent meaning and persistent uncertainties and ambiguities; bringing artistic, evocative, and visualization techniques to data analysis and presentations; and inviting outside-the-box, off-the-wall, and beyond-the-ken perspectives and interpretations.

Evaluative thinking forces clarity about the inquiry purpose, who it is for, with what intended uses, to be judged by what quality criteria; it involves being explicit about what criteria are being applied in framing inquiry questions, making design decisions, determining what constitutes *appropriate* methods, and selecting and following analytical processes and being aware of and articulating values, ethical considerations, contextual implications, strengths and weaknesses of the inquiry, and potential (or actual) misinterpretations, misuses, and misapplications. In contrast with the perspective of rigor as strict adherence to a standardized process, evaluative thinking emphasizes the importance of understanding the sufficiency of rigor relative to context and situational factors (Clarke, 2005; Patton, 2012a).

Inferential thinking involves examining the extent to which the evidence supports the conclusions reached. Inferential thinking can be deductive, inductive, or abductive—and often draws on and creatively integrates all three analytical processes—but at the core, it is a fierce examination of and allegiance to where the evidence leads.

> A rigorously conducted evaluation will be convincing as a presentation of evidence in support of an evaluation's conclusions and will presumably be more successful in withstanding scrutiny from critics. Rigor is multifaceted and relates to multiple dimensions of the evaluation. . . . The concept of rigor is understood and interpreted within the larger context of validity, which concerns the "soundness or trustworthiness of the inferences that are made from the results of the information gathering process" (Joint Committee on Standards for Educational Evaluation, 1994, p. 145). . . . There is relatively broad consensus that validity is a property of an inference, knowledge claim, or intended use, rather than a property either of a research or evaluation study from the study's findings. (Braverman, 2013, p. 101)

In reflecting on and writing about "what counts as credible evidence in applied research and evaluation practice," Sharon Rallis (2009), former president of the AEA and experienced qualitative researcher, emphasized rigorous reasoning: "I have come to see *a true scientist* [italics added], then, as one who puts forward her findings and the reasoning that led her to those findings for others to contest, modify, accept, or reject" (p. 171).

Practical thinking calls for assiduously integrating theory and practice, examining real-world implications of findings, inviting interpretations and applications from nonresearchers (e.g., community members, program staff, and participants) who can and will apply to the data what ordinary people refer to as "common sense"; and applying real-world criteria to interpreting the findings, criteria like understandability, meaningfulness, cost implications, and implications in addressing societal issues and problems.

What's at Stake?

> My words fly up, my thoughts remain below:
>
> Words without thoughts, never to heaven go.
>
> —William Shakespeare (1564–1616)
> The king in *Hamlet*

As I noted in Chapter 4, and is worth repeating here, philosopher Hannah Arendt (1968) concluded that to resist efforts by the powerful to deceive and control thinking, people need to practice thinking: "Experience

in thinking . . . can be won, like all experience in doing something, only through practice, through exercises" (p. 4).

Regardless of what one thinks of the U.S. invasion of Iraq to depose Saddam Hussein in 2003, both those who supported the war and those who opposed it ultimately agreed that the intelligence used to justify the invasion was deeply flawed and systematically distorted (U.S. Senate Select Committee on Intelligence, 2004). Under intense political pressure to show sufficient grounds for military action, those charged with analyzing and evaluating intelligence data began doing what is sometimes called cherry-picking or stove-piping—selecting and passing on only those data that support preconceived positions and ignoring or repressing all contrary evidence (Hersh, 2003; Tan, 2014; Zelik et al., 2007). The failure of the intelligence community to appropriately and accurately assess whether Iraq had weapons of mass destruction was not a function of poor data but of weak analysis, political manipulation of the analysis process, and a fundamental failure to think critically, creatively, evaluatively, and practically. The generation of the Rigor Attribute Model to support more rigorous intelligence analysis and restore credibility to the intelligence community focuses on *rigorous thinking* (Zelik et al., 2007; see Exhibit 9.5, pp. 675–677).

> Despite the etymological implication that to be rigorous is to "be stiff," expert information analysis processes often are not rigid in their application of a standard process, but rather, flexible and adaptive to highly dynamic environments. In information analysis, judgment of rigor reflects a relationship in the appropriateness of fit between analytic processes and contextual requirements. Thus, as supported by this and other research, rigor is more meaningfully viewed as an assessment of degree of sufficiency, rather than degree of adherence to an established analytic procedure. (Zelik, Patterson, & Woods, 2007, p. 1)

The phrase "degree of sufficiency" as a criterion for assessing rigor refers to an evaluation of the extent to which a multidimensional, multiperspectival, and critical thinking process was followed determinedly to yield conclusions that best fit the data, and therefore findings that are credible to and inspire confidence among those who must use the findings.

Bottom-Line Conclusion

Methods do not ensure rigor. A research design does not ensure rigor. Analytical techniques and procedures do not ensure rigor. Rigor resides in, depends on, and is manifest in *rigorous thinking*—about everything, including methods and analysis.

The thread that runs through this rumination is the importance of intellectual rigor. There are no simple formulas or clear-cut rules about how to do a credible, high-quality analysis. The task is to do one's best to make sense of things. A qualitative analyst returns to the data over and over again to see if the constructs, categories, interpretations, and explanations make sense—if they sufficiently reflect the nature of the phenomena studied. Creativity, intellectual rigor, perseverance, insight—these are the intangibles that go beyond the routine application of scientific procedures. It is worth quoting again Nobel prize–winning physicist Percy Bridgman: "There is no scientific method as such, but the vital feature of a scientist's procedure has been merely to do his utmost with his mind, *no holds barred*" (quoted in Waller, 2004, p. 106).

Varieties of and Concerns About Reactivity: How What We See and Do Affects What Is Seen and Done

> Nasrudin denied that he was a fisherman. From a passing tourist he had heard of something called philanthropy and, feeling transformed by what he had learned, he instantly adopted the moniker for himself. He explained to his fellow villagers: "When we see a problem that needs solving, it is wrong to just stand by and observe as scholars are wont to do. We must react. It is wrong to remain passive and detached in the face of need and noble to render help."
>
> "I am a philanthropist. Each day I strive to help fish that are drowning in the lake. I save them. I throw out my net and the fish rush in. I quickly put the many fish I've rescued on the dry ground, where they dance about in joy. But the dancing soon exhausts them and before long they cease to move. Alas, they dance themselves to death."
>
> "It is sad, but it is also wrong not to honor their struggle. So I take the dead fish to market where people contribute money to my effort to save more fish in exchange for my gifts to them of those fish who have lost the struggle. With the financial tokens of appreciation I receive for my charitable work, I purchase more nets so I can rescue more fish."
>
> —From Halcolm's *Chronicles of Lessons Learned: Teach a Man to Fish*

SIDEBAR

IN-DEPTH REFLEXIVITY: GUIDELINES FOR QUALITY IN AUTOBIOGRAPHICAL FORMS OF SELF-STUDY RESEARCH

- Autobiographical self-studies should ring true and enable connection.
- Self-studies should promote insight and interpretation.
- Autobiographical self-study research must engage history forthrightly, and the author must take an honest stand.
- Authentic voice is a necessary but not sufficient condition for the scholarly standing of a biographical self-study.
- The autobiographical self-study researcher has an ineluctable obligation to seek to improve the learning situation not only for the self but also for the other.
- Powerful autobiographical self-studies portray character development and include dramatic action: Something genuine is at stake in the story.
- Quality autobiographical self-studies attend carefully to persons in context or setting.
- Quality autobiographical self-studies offer fresh perspectives on established truths.
- To be scholarship, edited conversation or correspondence must not only have coherence and structure but that coherence and structure should also provide argumentation and convincing evidence.
- Interpretations made of self-study data should not only reveal but also interrogate the relationships, contradictions, and limits of the views presented (adapted from Bullough & Pinnegar, 2001, pp. 13–21).

Considering and Reporting Investigator Effects: Varieties of Reactivity

Reflectivity includes considering and reporting how your presence as an observer or evaluator may have affected what you observed. There are four primary ways in which the presence of an outside observer, or the fact that an evaluation is taking place, can affect, and possibly distort, the findings of a study, namely,

1. reactions of those in the setting (e.g., program participants and staff) to the presence of the qualitative fieldworker;
2. changes in you, the fieldworker (the measuring instrument), during the course of the data collection or analysis—that is, what has traditionally been called instrumentation effects in quantitative measurement;
3. the predispositions, selective perceptions, and/or biases you might bring to the inquiry that become evident to others during data collection; and
4. researcher incompetence (including lack of sufficient training or preparation).

Reactivity

> **All accounts produced by researchers must be interpreted within the context in which they were generated. Interpretations must examine, as carefully as possible, how the presence of the researcher, the context in which data were obtained, and so on shaped the data.**
>
> —Schwandt (2007, p. 256)

Problems of reactivity are well documented in the anthropological literature, which is one of the prime reasons why qualitative methodologists advocate long-term observations that permit an initial period during which observers and the people in the setting being observed get a chance to get used to each other. This increases trustworthiness, which supports credibility both within and outside the study setting.

> The credibility of your findings and interpretations depend upon your careful attention to establishing trustworthiness. . . . Time is a major factor in the acquisition of trustworthy data. Time at your research site, time spent interviewing, and time building sound relationships with respondents all contribute to trustworthy data. When a large amount of time is spent with your research participants, they less readily feign behavior or feel the need to do so; moreover, they are more likely to be frank and comprehensive about what they tell you. (Glesne, 1999, p. 151)

On the other hand, prolonged engagement may actually increase reactivity as the researcher becomes more a part of the setting and begins to affect what goes on through prolonged engagement. Thus, whatever the length of inquiry or method of data collection, researchers have an obligation to examine how their presence affects what goes on and what is observed.

> It is axiomatic that observers must record what they perceive to be their own reactive effects. They may treat this reactivity as bad and attempt to avoid it (which is impossible), or they may accept the fact that they will have a reactive effect and attempt to use it to advantage. . . . The reactive effect will be measured by daily

> field notes, perhaps by interviews in which the problem is pointedly inquired about, and also in daily observations. (Denzin, 1978b, p. 200)

Anxieties that surround an evaluation can exacerbate reactivity. The presence of an evaluator can affect how a program operates as well as its outcomes. The evaluator's presence may, for example, create a halo effect so that staff perform in an exemplary fashion and participants are motivated to "show off." On the other hand, the presence of the evaluator may create so much tension and anxiety that performances are below par. Some forms of program evaluation, especially "empowerment evaluation" and "intervention-oriented evaluation," (Patton, 2008, chap. 5) turn this traditional threat to validity into an asset by designing data collection to enhance achievement of the desired program outcomes. For example, at the simplest level, the observation that "what gets measured gets done" suggests the power of data collection to affect outcomes attainment. A leadership program, for example, that includes in-depth interviewing and participant journal writing as ongoing forms of evaluation data collection may find that participating in the interviewing and writing reflectively have effects on participants' learning and program outcomes. Likewise, a community-based AIDS awareness intervention can be enhanced by having community participants actively engaged in identifying and doing case studies of critical community incidents. In short, a variety of reactive responses are possible, some that support program processes, some that interfere, and many that have implications for interpreting findings. Thus, the evaluator has a responsibility to think about the problem, make a decision about how to handle it in the field, attempt to monitor evaluator/observer effects, and reflect on how reactivities may have affected the findings.

Evaluator effects can be overrated, particularly by evaluators. There is more than a slight touch of self-importance in some concerns about reactivity. Lillian Weber, director of the Workshop Center for Open Education, City College School of Education, New York, once set me straight on this issue, and I pass her wisdom on to my colleagues. In doing observations of open classrooms, I was concerned that my presence, particularly the way kids flocked around me as soon as I entered the classroom, was distorting the evaluation to the point where it was impossible to do good observations. Lillian laughed and suggested to me that what I was experiencing was the way those classrooms actually were. She went on to note that this was common among visitors to schools; they were always concerned that the teacher, knowing visitors were coming, whipped the kids into shape for those visitors. She suggested that under the best of circumstances a teacher might get kids to move out of habitual patterns into some model mode of behavior for as much as 10 or 15 minutes but that, habitual patterns being what they are, the kids would rapidly revert to normal behaviors and whatever artificiality might have been introduced by the presence of the visitor would likely become apparent.

Evaluators and researchers should strive to neither overestimate nor underestimate their effects but to take seriously their responsibility to describe and study what those effects are.

Effects on the Inquirer of Being Engaged in the Inquiry

A second form of reactivity arises from the possibility that the researcher or evaluator changes during the course of the inquiry. In Chapter 7, on interviewing, I offered several examples of this, including how in a study of child sexual abuse, those involved were deeply affected by what they heard. One of the ways this sometimes happens in anthropological research is when participant observers "go native" and become absorbed into the local culture. The epitome of this in a short-term observation is the legendary story of the student observers who became converted to Christianity while observing a Billy Graham evangelical crusade (Lang & Lang, 1960). Evaluators sometimes become personally involved with program participants or staff and therefore lose their sensitivity to the full range of events occurring in the setting.

Johnson (1975) and Glazer (1972) have reflected on how they and others have been changed by doing field research. The consensus of advice on how to deal with the problem of changes in observers as a result of involvement in research is similar to advice about how to deal with the reactive effects created by the presence of observers.

> It is central to the method of participant observation that changes will occur in the observer; the important point, of course, is to record these changes. Field notes, introspection, and conversations with informants and colleagues provide the major means of measuring this dimension, . . . for to be insensitive to shifts in one's own attitudes opens the way for placing naive interpretations on the complex set of events under analysis. (Denzin, 1978b, p. 200)

Inquirer–Selective Perception and Predispositions

The third concern about inquirer effects related to credibility has to do with the extent to which the predispositions or biases of the inquirer may affect data

analysis and interpretations. This issue carries mixed messages because, on the one hand, rigorous data collection and analytical procedures, like triangulation, are aimed at substantiating the credibility of the findings and minimizing inquirer biases and, on the other, the interpretative and constructivist perspectives remind us that data from and about humans inevitably represent some degree of perspective rather than absolute truth. Getting close enough to the situation observed to experience it firsthand means that researchers can learn from their experiences, thereby generating personal insights; but that closeness makes their objectivity suspect. "For social scientists to refuse to treat their own behavior as data from which one can learn is really tragic" (Scriven, 1972a, p. 99). In effect, all of the procedures for validating and verifying analysis that have been presented in this chapter are aimed at reducing distortions introduced by inquirer predisposition. Still, people who use different criteria in determining evidential credibility will come at this issue from different stances and end up with different conclusions.

Consider the interviewing stance of *emphatic neutrality* introduced in Chapter 2 and elaborated in Chapter 7. An emphatically neutral inquirer will be perceived as caring about and interested in the people being studied but neutral about the content of what they reveal. House (1977) balances the caring, interested stance against independence and impartiality for evaluators, a stance that also applies to those working according to the standards of traditional science.

> The evaluator must be seen as caring, as interested, as responsive to the relevant arguments. He must be impartial rather than simply objective. The impartiality of the evaluator must be seen as that of an actor in events, one who is responsive to the appropriate arguments but in whom the contending forces are balanced rather than non-existent. The evaluator must be seen as not having previously decided in favor of one position or the other. (pp. 45–46)

But neutrality and impartiality are not easy stances to achieve. Denzin (1989b) cites a number of scholars who have concluded, as he does, that every researcher brings preconceptions and interpretations to the problem being studied, regardless of the methods used.

> All researchers take sides, or are partisans for one point of view or another. Value-free interpretive research is impossible. This is the case because every researcher brings preconceptions and interpretations to the problem being studied. The term *hermeneutical circle or situation* refers to this basic fact of research. All scholars are caught in the circle of interpretation. They can never be free of the hermeneutical situation. This means that scholars must state beforehand their prior interpretations of the phenomenon being investigated. Unless these meanings and values are clarified, their effects on subsequent interpretations remain clouded and often misunderstood. (p. 23)

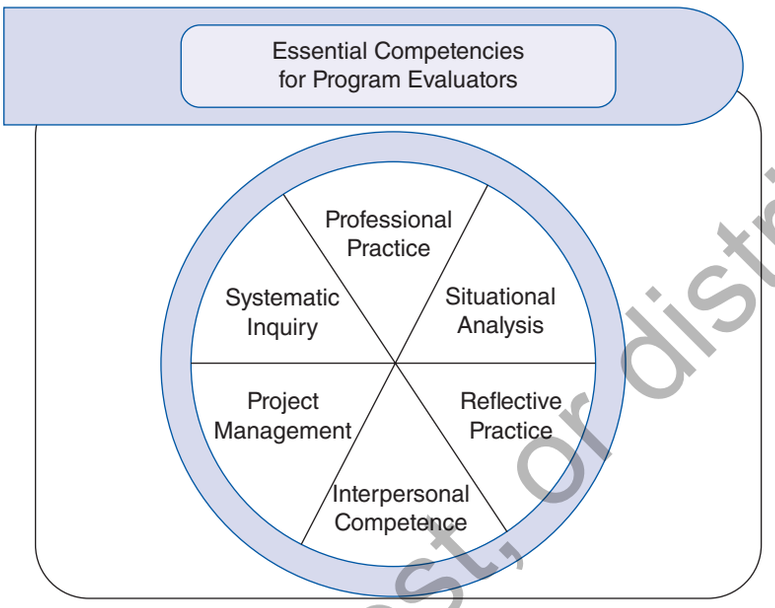
Earlier I presented seven sets of criteria for judging the quality of qualitative inquiry (Exhibit 9.7, pp. 680–681). Those varying and competing frameworks offer different perspectives on how inquirers should deal with concerns about bias. Neutrality and impartiality are expected when qualitative work is being judged by traditional scientific criteria or by evaluation standards, thus the source of House's (1977) admonition quoted above. In contrast, constructivist analysts are expected to deal with these issues through conscious and committed reflexivity—entering the *hermeneutical circle of interpretation* and therein reflecting on and analyzing how their perspective interacts with the perspectives they encounter. Artistic inquirers often deal with issues of how they personally relate to their work by invoking aesthetic criteria: Judge the work on its *artistic merits*. Participatory and collaborative inquiries encourage the formation of meaningful and trusting relationships between researchers and those participating in the inquiry. When critical change criteria are applied in judging reactivity, the issue becomes whether, how, and to what extent the inquiry furthered the cause or enhanced the well-being of those involved and studied; neutrality is eschewed in favor of explicitly using the inquiry process to facilitate change, or at least illuminate the conditions needed for change.

Inquirer Competence

Concerns about the extent to which the inquirer's findings can be trusted—that is, trustworthiness—can be understood as one dimension of perceived methodological rigor. But ultimately, for better or worse, the trustworthiness of the data is tied directly to the trustworthiness of those who collect and analyze the data—and their demonstrated competence. Competence is demonstrated by using the verification and validation procedures necessary to establish the quality of analysis and thereby building a "track record" of quality work. As Exhibit 9.10 shows, inquirer competence includes not just systematic inquiry knowledge and skill but also interpersonal competence, reflective practice skills, situational analysis, professional practice competence, and project management. This array of competencies is being acknowledged and certified by professional evaluation associations around the world (King & Podems, 2014; Podems, 2014a). Consistent

with the overall message of this chapter, especially my MQP Rumination on avoiding research rigor mortis, thinking skills also need ongoing development. An excellent resource in that regard is the *Critical Evaluation Skills Toolkit* (Crebert, Patrick, Cragnolini, Smith, Worsfold, & Webb, 2011).

EXHIBIT 9.10 The Multiple Dimensions of Program Evaluator Competence



SOURCES: Ghere, King, Stevahn, and Minnema (2006) and King, Stevahn, Ghere, and Minnema (2001).

The principle for dealing with inquirer competence is this: Don't wait to be asked. Anticipate competence as an issue. Address the issue of competence proactively, explicitly, and multidimensionally. With quantitative methods, validity and reliability reside in tools, instruments, design parameters, and procedures. In qualitative inquiry, the competency stakes are greater because the inquirer is the instrument. Trustworthiness and authenticity are functions of systematic inquiry procedures, interpersonal (relational) dynamics in the field, and competency to engage in the challenges and deal with the ambiguities of qualitative inquiry.

Review: The Credibility of the Inquirer

Because the researcher is the instrument in qualitative inquiry, the credibility of the inquirer is central to the credibility of the study. Exhibit 9.11 on the next page summarizes the issues that arise in establishing and judging the credibility of the inquirer.

EXHIBIT 9.11 The Credibility of the Inquirer: Issues and Solutions

CREDIBILITY CONCERN	WHAT'S THE ISSUE?	WAYS TO ADDRESS THE ISSUES AND ENHANCE CREDIBILITY	EXAMPLES FROM NORA MURPHY'S (2014) STUDY OF HOMELESS YOUTH
1. Who did the study? Who is the inquirer?	Because the researcher is the instrument in qualitative inquiry, who did the work and carried out the analysis matters.	The methodology section of the report should present not only the usual design and data collection details and rationale but also description of the inquirer's relevant experiences, training, perspective, competence, and purpose.	A University of Minnesota doctoral student on the staff of the Minnesota Evaluation Studies Institute, who has completed doctoral studies in evaluation theory, methods, and practice and is being supervised by experienced and knowledgeable researchers and evaluators, did the study.
2. Reflexivity	How has the inquirer's background and perspective affected the findings?	Reflexivity goes beyond reporting background, experience, and training; it involves reflecting on and reporting your reflexive process and the answers to reflexive questions: How do you know what you know? What shapes and has shaped your perspective? (See Exhibit 2.5, p. 72.)	"I am a constructivist working from a systems perspective. My evaluation approach is *utilization focused*, and I'm using *developmental evaluation* because it fits the dynamics, complexities, and developmental nature of the initiative being evaluated. I have worked as a teacher and program staff member with disadvantaged youth."
3. Potential inquirer bias	How might the findings be a function of the inquirer's selective perception, predispositions, and bias? What steps have been taken to deal with potential bias?	*Options (not mutually exclusive)* a. Acknowledge potential sources of bias: What brought you to this inquiry? Why do you care about what you're studying? What are the implications of caring? How do you deal with concerns about bias (which you acknowledge as legitimate)? b. Acknowledge your perspective and present it as a strength: "I am a constructivist and view getting close enough to people to experience empathy as a strength of in-depth fieldwork and interviewing." c. Describe your process for surfacing and setting aside any preconceptions (e.g., *epoche* in phenomenological analysis). d. Subject your analysis to independent review (analyst triangulation, peer review, or external audit).	"I care about homeless youth and believe they deserve an opportunity to move on in their life's journey past their period of homelessness. I believe in and subscribe to the values and principles expressed by the programs participating in the evaluation. I want to help them better elucidate and implement those principles. I also want evaluation to be a vehicle for giving voice to homeless young people, to honor their stories, and help them articulate what they've experienced." The study is being done collaboratively with six youth-serving agencies, which have monitored the appropriateness and integrity of the data collection and analysis. The study is supervised by experienced researchers and evaluators who monitor the integrity of the methods and analysis.

CREDIBILITY CONCERN	WHAT'S THE ISSUE?	WAYS TO ADDRESS THE ISSUES AND ENHANCE CREDIBILITY	EXAMPLES FROM NORA MURPHY'S (2014) STUDY OF HOMELESS YOUTH
4. Reactivity	How has the inquiry affected the people in the setting studied?	*Options (not mutually exclusive)* Demonstrate awareness of the issue and take it seriously: a. Keep field notes on your observations about how your presence may have, or actually did appear, to affect things. Describe effects and their implications for your findings. b. Gather data about reactions; ask key informants how your presence affected the setting observed and people interviewed.	a. Youth interviewed were compensated for their time. They reviewed and approved the case studies created from their interview transcripts. They expressed appreciation for the opportunity to tell their stories. b. The six participating agencies report having strengthened their collaboration with each other as a result of being part of this study (which was the intent). They reported learning from the experience, feeling that their work and approach was validated, and are using the findings for staff development in their agencies.
5. Effects on the inquirer of involvement in the inquiry	How were you affected or changed by engaging in this inquiry?	Reflect on and report what you've seen and how it has affected you. Acknowledge emotional responses and any actions taken. Acknowledge that as the research instrument, you are also a human being—and report honestly your human responses.	"This study took a personal toll. There were times when I went home and cried. I felt guilt that I could not help them more and fear that I was exploiting them. What helped me was that listening seemed to help them. I still carry some of the sadness that I experienced as I sat with the youth and their telling of their lives, but I also carry the hope that I felt when I experienced their optimism and their strength."
6. Competence	How can I, the reader and user of your findings, be assured of your competence to undertake this inquiry?	Acknowledge the importance of competence and its multiple dimensions (see Exhibit 9.10), and report on your competence in these areas.	The entire collaboration engaged in reflective practice together. Confidentiality, rapport, and trust were essential in interviewing the youth. Sensitivity to race, gender orientation, and the trauma experienced by homeless youth were monitored by the participating agencies. The methods section of the study addresses these and related issues in depth.

MODULE

81 Generalizations, Extrapolations, Transferability, Principles, and Lessons Learned

> **The trouble with generalizations is that they don't apply to particulars.**
>
> —Lincoln and Guba (1985, p. 110)

Credibility and utility are linked. What can one do with qualitative findings? The results illuminate a particular situation or small number of cases. But can qualitative findings be generalized? Here, again, different qualitative frameworks based on different criteria offer different answers. The traditional scientific research criteria include generalizability. Constructivist criteria, in contrast, emphasize particularity; constructivists generally eschew, and are skeptical about, generalizability. They offer *extrapolations* and *transferability* instead. So let's see if we can sort out these different perspectives and their implications.

Purposeful Sampling and Generalizability

Chapter 5 discussed the logic and value of purposeful sampling with small but carefully selected information-rich cases. Certain kinds of small samples and qualitative studies are designed for generalizability and broader relevance: a critical case, an index case, a causal pathway sample, a positive deviance case, and a qualitative synthesis review are examples (see Exhibit 5.8, pp. 266–272). Other sampling strategies, for example, outlier cases (exemplars of excellence or failure), a high-impact case, sensitizing concept exemplars, and principles-focused sampling, aim to yield insights about principles that might be adapted for application elsewhere. In short, the conditions for, possibility of, and relative importance attached to generalizability are determined at the design stage. *To review:* Purpose drives design. Design drives data collection. Data drive analysis. Purpose, design, data, and analysis, in combination, determine generalizability.

Principles of Generalizability

Shadish (1995a) has made the case that certain core principles of generalization apply to both experiments and ethnographies (or qualitative methods generally). Both experiments and case studies share the problem of being highly localized. Findings from a study, experimental or naturalistic in design, can be generalized according to five principles:

> 1. ***The principle of proximal similarity:*** We generalize most confidently to applications where treatments, settings, populations, outcomes, and times are most similar to those in the original research. . . .
>
> 2. ***The principle of heterogeneity of irrelevancies:*** We generalize most confidently when a research finding continues to hold over variations in persons, settings, treatments, outcome measures, and times that are presumed to be conceptually irrelevant. The strategy here is identifying irrelevancies, and where possible including a diverse array of them in the research so as to demonstrate generalization over them. . . .
>
> 3. ***The principle of discriminant validity:*** We generalize most confidently when we can show that it is the target construct, and not something else, that is necessary to produce a research finding. . . .
>
> 4. ***The principle of empirical interpolation and extrapolation:*** We generalize most confidently when we can specify the range of persons, settings, treatments, outcomes, and times over which the finding holds more strongly, less strongly, or not all. The strategy here is empirical exploration of the existing range of instances to discover how that range might generate variability in the finding for instances not studied. . . .
>
> 5. ***The principle of explanation:*** We generalize most confidently when we can specify completely and exactly (a) which parts of one variable (b) are related to which parts of another variable (c) through which mediating processes (d) with which salient interactions, for then we can transfer only those essential components to the new application to which we wish to generalize. The strategy here is breaking down the finding into component parts and processes so as to identify the essential ones. (pp. 424–426)

Generalizability Versus Contextual Particularity

Deep philosophical and epistemological issues are embedded in concerns about generalizing. What's desirable or hoped for in science (generalizations across

SIDEBAR

CULTURAL LIMITS ON GENERALIZABILITY

Psychological experiments have been used to study how people react to things like negotiating rewards and perceptions of whether two lines are of equal length when phony participants in the experiment say that the shorter line is really longer. The results of most such laboratory research have been interpreted as showing "evolved psychological traits common to all humans" (Watters, 2013, p. 1). However, when such experiments are repeated in other cultures, the ways in which Americans respond can be quite different from how nonliterate peoples respond. What were once thought to be tests of basic perception (how the brain works) have turned out to be culturally determined. Social scientists had assumed that lab experiments studied

> the human mind stripped of culture, [that] the human brain is genetically comparable around the globe, it was agreed, so human hardwiring for much behavior, perception, and cognition should be similarly universal. No need, in that case, to look beyond the convenient population of undergraduates for test subjects. A 2008 survey of the top six psychology journals dramatically shows how common that assumption was: more than 96 percent of the subjects tested in psychological studies from 2003 to 2007 were Westerners—with nearly 70 percent from the United States alone. Put another way: 96 percent of human subjects in these studies came from countries that represent only 12 percent of the world's population. (Watters, 2013, p. 1)

Cross-cultural research is now revealing that

> the mind's capacity to mold itself to cultural and environmental settings was far greater than had been assumed. The most interesting thing about cultures may not be in the observable things they do—the rituals, eating preferences, codes of behavior, and the like—but in the way they mold our most fundamental conscious and unconscious thinking and perception. (Watters, 2013, p. 1)

Moreover, the experiments done on American undergraduate students may be especially prone to inappropriate overgeneralizations.

> It is not just our Western habits and cultural preferences that are different from the rest of the world, it appears. The very way we think about ourselves and others—and even the way we perceive reality—makes us distinct from other humans on the planet, not to mention from the vast majority of our ancestors. Among Westerners, the data showed that Americans were often the most unusual, leading the researchers to conclude that "American participants are exceptional even within the unusual population of Westerners—outliers among outliers."
>
> Given the data, they concluded that social scientists could not possibly have picked a worse population [American undergraduate students)] from which to draw broad generalizations. Researchers had been doing the equivalent of studying penguins while believing that they were learning insights applicable to all birds. (Watters, 2013, p. 1)

time and space) runs into real-world considerations about what's possible. Lee J. Cronbach (1975), one of the major figures in psychometrics and research methodology in the twentieth century, devoted considerable attention to the issue of generalizations. He concluded that social phenomena are too variable and context bound to permit very significant empirical generalizations. He compared generalizations in natural sciences with what was likely to be possible in behavioral and social sciences. His conclusion was that "generalizations decay. At one time a conclusion describes the existing situation well, at a later time it accounts for rather little variance, and ultimately is valid only as history" (p. 122).

Cronbach (1975) offers an alternative to generalizing that constitutes excellent advice for the qualitative analyst:

> Instead of making generalization the ruling consideration in our research, I suggest that we reverse our priorities. An observer collecting data in a particular situation is in a position to appraise a practice or proposition in that setting, observing effects in context. In trying to describe and account for what happened, he will give attention to whatever variables were controlled, but he will give equally careful attention to uncontrolled conditions, to personal characteristics, and to events that occurred during treatment and measurement. As he goes from situation to situation, his first task is to describe and interpret the effect anew in each locale, perhaps taking into account factors unique to that locale or series of events. . . . When we give proper weight to local conditions, any generalization is a working hypothesis, not a conclusion. (pp. 124–125)

Robert Stake (1978, 1995, 2000, 2006, 2010), master of the case study, concurs with Cronbach that the first priority is to do justice to the specific case, to do a good job of "particularization" before looking for patterns across cases. He quotes William Blake on the subject:

> To generalize is to be an idiot. To particularize is the lone distinction of merit. General knowledges are those that idiots possess.

Stake (1978) continues,

> Generalization may not be all that despicable, but particularization does deserve praise. To know particulars fleetingly, of course, is to know next to nothing. What becomes useful understanding is a full and thorough knowledge of the particular, recognizing it also in new and foreign contexts. That knowledge is a form of generalization too, not scientific induction but naturalistic generalization, arrived at by recognizing the similarities of objects and issues in and out of context and by sensing the natural covariations of happenings. To generalize this way is to be both intuitive and empirical, and not idiotic. (p. 6)

Stake (2000) extends *naturalistic generalizations* to include the kind of learning that readers take from their encounters with specific case studies. The "vicarious experience" that comes from reading a rich case account can contribute to the social construction of knowledge, which, in a cumulative sense, builds general, if not necessarily generalizable, knowledge.

> Readers assimilate certain descriptions and assertions into memory. When researcher's narrative provides opportunity for vicarious experience, readers extend their memories of happenings. Naturalistic, ethnographic case materials, to some extent, parallel actual experience, feeding into the most fundamental processes of awareness and understanding . . . [to permit] *naturalistic generalizations*. The reader comes to know some things told, as if he or she had experienced it. Enduring meanings come from encounter, and are modified and reinforced by repeated encounter.
>
> In life itself, this occurs seldom to the individual alone but in the presence of others. In a social process, together they bend, spin, consolidate, and enrich their understandings. We come to know what has happened partly in terms of what others reveal as their experience. The case researcher emerges from one social experience, the observation, to choreograph another, the report. Knowledge is socially constructed, so we constructivists believe, and, in their experiential and contextual accounts, case study researchers assist readers in the construction of knowledge. (p. 442)

Guba (1978) considered three alternative positions that might be taken in regard to the generalizability of naturalistic inquiry findings:

> 1. Generalizability is a chimera; it is impossible to generalize in a scientific sense at all. . . .
>
> 2. Generalizability continues to be important, and efforts should be made to meet normal scientific criteria that pertain to it. . . .
>
> 3. Generalizability is a fragile concept whose meaning is ambiguous and whose power is variable. (pp. 68–70)

Having reviewed these three positions, Guba (1978) proposed a resolution that recognizes the diminished value and changed meaning of generalizations and echoes Cronbach's emphasis, cited above, on treating conclusions as hypotheses for future applicability and testing rather than as definitive.

> The evaluator should do what he can to establish the generalizability of his findings. . . . Often naturalistic inquiry can establish at least the "limiting cases" relevant to a given situation. But in the spirit of naturalistic inquiry he should regard each possible generalization only as a working hypothesis, to be tested again in the next encounter and again in the encounter after that. For the naturalistic inquiry evaluator, premature closure is a cardinal sin, and tolerance of ambiguity a virtue. (p. 70)

Guba and Lincoln (1981) emphasized appreciation of and attention to context as a natural limit to naturalistic generalizations. They ask, "What can a generalization be except an assertion that is context free? [Yet] *it is virtually impossible to imagine any human behavior that is not heavily mediated by the context in which it occurs*" (p. 62). They proposed substituting the concepts "transferability" and "fittingness" for generalization when dealing with qualitative findings:

> The degree of *transferability* is a direct function of the *similarity* between the two contexts, what we shall call "*fittingness*." Fittingness is defined as degree of congruence between sending and receiving contexts. If context A and context B are "sufficiently" congruent, then working hypotheses from the sending originating context may be applicable in the receiving context. (Lincoln & Guba, 1985, p. 124)

Cronbach (1980) offered a middle ground in the debate over generalizability. He found little value in experimental designs that are so focused on carefully controlling cause and effect (internal validity) that the findings are largely irrelevant beyond that highly controlled experimental situation (external validity). On the other hand, he was equally concerned about entirely

idiosyncratic case studies that yield little of use beyond the case study setting. He was also skeptical that highly specific empirical findings would be meaningful under new conditions. He suggested instead that designs balance depth and breadth, realism and control so as to permit reasonable "extrapolation" (pp. 231–235).

SIDEBAR

TESTING THEORY FROM A PURPOSEFUL SAMPLE OF QUALITATIVE CASES TO GENERALIZE: A CLASSIC CASE EXAMPLE

Sociologist Alfred Lindesmith (1905–1991), Indiana University, wanted to test his theory about addiction to opiate drugs. The theory posited that people became addicted to opium, morphine, or heroin when they took the drug often enough and in sufficient quantity to develop physical withdrawal. But Lindesmith had observed that people become habituated to opiates in a hospital when medicated for pain and manifest junkie behavior of compulsively searching for drugs at almost any cost after hospitalization. He hypothesized that two other things had to happen: Having become habituated, the potential addict now had to (1) stop using drugs and experience the painful withdrawal symptoms that resulted and (2) consciously connect withdrawal distress with ceasing drug use, a connection not everyone made. Junkies, unlike former hospital patients, then had to act on that realization and take more drugs to relieve the symptoms. Those steps, taken together and taken repeatedly, create the compulsive activity that is addiction.

A well-known statistician criticized Lindesmith's sample because he had generalized to a large population (all the addicts in the United States or in the world) from a small, purposefully selected sample rather than studying a random sample. Lindesmith replied that the purpose of random sampling was to ensure that every case had a known probability of being drawn for a sample and that researchers randomize to permit generalizations about distributions of some phenomenon in a population and in subgroups in a population. But, he argued, random sampling was irrelevant to his research on addicts because he was interested not in distributions but in a universal process—how one became and remained an addict. He didn't want to know the probability that any particular case would be chosen for his sample. He wanted to maximize the probability of finding a negative case so as all the better to test the theory. Not finding disconfirming cases strengthened his confidence in generalizing his findings.

—Adapted from Becker (1998, pp. 86–87)

Extrapolation

Unlike the usual meaning of the term *generalization*, an *extrapolation* clearly connotes that one has gone beyond the narrow confines of the data to *think about other applications of the findings*. Extrapolations are modest speculations on the likely applicability of findings other situations under similar, but not identical, conditions. Extrapolations are logical, thoughtful, case derived and problem oriented rather than statistical and probabilistic.

Distinguished methodologist Thomas D. Cook (2014) has explained the nature and significance of extrapolation.

> Informing future policy decisions also requires justified procedures for extrapolating past findings to future periods when the populations of treatment providers and recipients might be different, when adaptations of a previously studied treatment might be required, when a novel outcome is targeted, when the application might be to situations different from earlier, and when other factors affecting the outcome are novel too. We call this the *extrapolation function* since inferences are required about populations and categories that are now in some ways different from the sampled study particulars. Sampling theory cannot even pretend to deal with the framing of causal generalization as extrapolation since the emphasis is on taking observed causal findings and projecting them beyond the observed sampling specifics.
>
> We argue here that both representation and extrapolation are part of a broad and useful understanding of external validity; that each has been quite neglected in the past relative to internal validity—namely, whether the link between manipulated treatments and observed effects is plausibly causal; that few practical methods exist for validly representing the populations and other constructs sampled in the existing literature; and that even fewer such methods exist for extrapolation. Yet, causal extrapolation is more important for the policy sciences, I argue, than is causal representation. (p. 527)

Extrapolations can be particularly useful when based on information-rich samples and designs—that is, studies that produce relevant information carefully targeted to specific concerns about both the present and the future. Users of evaluation, for example, will usually expect evaluators to thoughtfully extrapolate from their findings in the sense of pointing out *lessons learned* and potential applications to future efforts. Sampling strategies in qualitative evaluations can be planned with the stakeholders' desire for extrapolation in mind.

High-Quality Lessons Learned

The notion of identifying and articulating "lessons learned" has become popular as a way of extracting useful and actionable knowledge from cross-case analyses. Rather than being stated in the form of traditional scientific empirical generalizations, lessons learned take the form of *principles of practice* that must be adapted to particular settings in which the principle is to be applied. For example, a lesson learned from research on evaluation use is that evaluation use will likely be enhanced by designing an evaluation to answer the focused questions of specific primary intended users (Cousins & Bourgeois, 2014; Patton, 2008).

Ricardo Millett, former Director of Evaluation at the W. K. Kellogg Foundation, and I analyzed the lessons-learned sections of grantee evaluation reports. What we found was massive confusion and inconsistency. Listed under the heading "lessons" were findings, opinions, ideas, visions, and recommendations—but seldom lessons. Exhibit 9.12 provides examples of what we found.

EXHIBIT 9.12 Confusion About What Constitutes *a Lesson Learned*

A lesson, in the context of extracting useable knowledge from findings, takes the form of an *if . . . then* proposition that provides direction for future action in the real world.

Lesson about evaluation use. If you actively involve intended users in designing an evaluation to ensure its relevance, they are more likely to be interested in and actually use the findings.

This lesson meets two criteria: (1) it is based on evidence from studies of evaluation use (Cousins & Bourgeois, 2014; Patton, 2008) and (2) it provides guidance for future action (an extrapolation from past evidentiary patterns to future desired outcomes). A lesson provides guidance, but it is different from a law, a recipe, or a theoretical proposition.

A physical law. If you heat water to 100 degrees Celsius at sea level, it will boil.

A recipe. Place a cup of oats in two cups of water, add a pinch of salt, and boil for five minutes. Remove from heat, and leave covered for two minutes. It is then ready to serve.

A theoretical proposition. It describes how the world works, as with *natural selection*: If a mutation provides a reproductive advantage that is heritable, over many generations that trait will become dominant in the population.

Using the definition of lesson and these distinctions, here is a sample of statements from evaluation reports illustrating confusion about what constitutes a lesson—and a lesson learned.

STATEMENT REPORTED UNDER THE HEADING "LESSONS" IN EVALUATION REPORTS	WHAT THE STATEMENTS ACTUALLY ARE
1. "Students whose parents helped them with homework got higher grades than those who did not get such help at home."	This is *a finding.* The lesson remains implicit and unexpressed.
2. "One size doesn't fit all."	This is *a conclusion* (based on findings that different people in a program wanted and needed different things), but the conditions to which this conclusion applies (the "if" statement) and what will result (the "then" statement) are implicit.
3. "There are no workarounds powerful enough to compensate for a failing educational system."	This is *an opinion* based on negative findings from an evaluation about a single program. It is a gross overgeneralization born of frustration and skepticism, but it lacks both supporting evidence and guidance about what to do in any applicable and useful manner.
4. "Be sure to provide daycare when you hold community meetings."	This is *a recommendation.* It prescribes a quite specific action, but both the basis for the recommendation and the outcome that will follow its implementation are implicit.

STATEMENT REPORTED UNDER THE HEADING "LESSONS" IN EVALUATION REPORTS	WHAT THE STATEMENTS ACTUALLY ARE
5. "Be prepared: By failing to prepare, you are preparing to fail."	This is *an aphorism*, no doubt wise, and certainly oft cited since published by Ben Franklin and adopted by the Boy Scouts, but reporting it as a central lesson in an evaluation report might at least include some acknowledgment that the observation has a long and distinguished history. (The report in which this appeared had nine others of like sentiment and no lessons original to or grounded in the actual evaluation done.)
6. "Lesson learned: Take time to do reflective practice."	This, too, is *a recommendation*, but it invites introduction of a useful distinction between "lessons" and "lessons learned." A lesson is a cognitive insight or understanding that if you do a certain thing, a certain result is likely to follow. A lesson is not "learned" until it is put into practice (behavioral change).
7. "We will never stop working to make the world a better place."	This is a *visionary promise*, an organizational commitment, and an inspirational reassurance to actual or potential funders. It is not a lesson.
8. If you want to formulate a meaningful and useful lesson that provides guidance for future action, then learn what a lesson is (as distinct from a finding, conclusion, opinion, recommendation, aphorism, or vision).	*That's a lesson.* If you put that lesson into action, you will have *a lesson learned.*

High-Quality Lessons

As we looked at examples of "lessons" listed in a variety of evaluation reports, it became clear that the label was being applied to any kind of insight, evidentially based or not. We began thinking about what would constitute "high-quality lessons" and decided that one's confidence in the transferability or extrapolated relevance of a supposed lesson would increase to the extent towhich it was supported by multiple sources and types of learnings (triangulation). Exhibit 9.13 on the next page presents a list of kinds of evidence that could be accumulated to support a proposed lesson, making it more worthy of application and adaptation to new settings if it has triangulated support from a variety of perspectives and data sources. Questions for generating lessons learned are also listed. Thus, for example, the lesson that designing an evaluation to answer the focused questions of specific primary intended users enhances evaluation use is supported by research on use, theories about diffusion of innovation and change, practitioner wisdom, cross-case analyses of use, the profession's articulation of standards, and expert testimony. *High-quality lessons*, then, constitute guidance extrapolated from multiple sources and independently triangulated to increase transferability as cumulative knowledge and working hypotheses that can be adapted and applied to new situations. This is a form of pragmatic utilitarian generalizability, if you will. The pragmatic bias in this approach reflects the wisdom of Samuel Johnson: "As gold which he cannot spend will make no man rich, so knowledge which he cannot apply will make no man wise."

Principles

Principles are lessons expressed more generically, taken to a higher level of generalizability, and stated in a more direct and less contingent manner.

Lesson about evaluation use: If you actively involve intended users in designing an evaluation to ensure its relevance, they are more likely to be interested in and actually use the findings.

Principle to enhance evaluation use: Form and nurture a relationship with primary intended users built around their information needs and intended uses of the evaluation.

Principles are built from lessons that are based on evidence about how to accomplish some desired result. Qualitative inquiry is an especially productive way

EXHIBIT 9.13 High-Quality Lessons Learned

High-quality lessons learned. Knowledge that can be applied to future action and derived from multiple sources of evidence (triangulation)

1. Evaluation findings—patterns across programs
2. Basic and applied research
3. Practice wisdom and experience of practitioners
4. Experiences reported by program participants/clients/intended beneficiaries
5. Expert opinion
6. Cross-disciplinary findings and patterns

The idea is that the greater the number and quality of supporting sources for a "lesson," the more rigorous the supporting evidence, and the greater the *triangulation of supporting sources*, the more confidence one has in the significance and meaningfulness of the lesson. Lessons promulgated with only one type of supporting evidence would be considered a "lessons" hypothesis. Nested within and cross-referenced to lessons should be the actual cases from which practice wisdom and evaluation findings have been drawn. A critical principle here is to maintain the contextual frame for lessons—that is, to keep lessons grounded in their context. For ongoing learning, the trick is to follow future-supposed applications of lessons to test their wisdom and relevance over time in action in new settings. If implemented and validated, they become *high-quality lessons learned*.

Questions for Generating High-Quality Lessons Learned

1. What is meant by a "*lesson*"?
2. What is meant by "*learned*"?
3. By whom was the lesson learned?
4. What's the evidence supporting each lesson?
5. What's the evidence the lesson was learned?
6. What are the contextual boundaries around the lesson (i.e., under what conditions does it apply)?
7. Is the lesson specific, substantive, and meaningful enough to guide practice in some concrete way?
8. Who else is likely to care about this lesson?
9. What evidence will they want to see?
10. How does this lesson connect with other "lessons"?

to generate lessons and principles precisely because purposeful sampling of information-rich cases, systematically and diligently analyzed, yields rich, contextually sensitive findings. This combination of qualitative elements constitutes the intellectual farming system from which nutritious lessons and principles grow and thrive. I have discussed principles-focused qualitative inquiry throughout this book.

- ***Chapter 1:*** Examples of principles as both a focus of inquiry (Paris Declaration Principle for Development Aid, p. 10) and the result of comparative case study analysis (principles that distinguish great from good organizations, Collins, 2001a; adaptive from nonadaptive companies, Collins & Hansen, 2011)
- ***Chapter 2:*** Strategic principles for qualitative inquiry (Exhibit 2.1, pp. 46–47)
- ***Chapter 3:*** Principles that undergird and guide various theoretical perspectives: constructivism, hermeneutics, pragmatism
- ***Chapter 4:*** Practical qualitative inquiry principles to get actionable answers (Exhibit 4.1, pp. 172–173); principles of fully participatory and genuinely collaborative inquiry (p. 222); and principles-focused evaluation (p. 194)
- ***Chapter 5:*** Principles-focused purposeful sampling (p. 292)
- ***Chapter 6:*** Principles for engaging in qualitative fieldwork (pp. 415–416)
- ***Chapter 7:*** Ten interview principles and skills (Exhibit 7.2, p. 428)
- ***Chapter 8:*** A principles-focused evaluation report (pp. 627–528)
- ***Chapter 9:*** Rigor attribute analysis principles (pp. 675–676)

How to Extract Credible and Useful Principles: A Case Example

> ***Scaling Up Excellence*** **tackles a challenge that confronts every leader and organization—spreading constructive beliefs and behavior from the few to the many. This book shows what it takes to build and uncover pockets of exemplary performance, spread those splendid deeds, and as an organization grows bigger and older—rather than slipping toward mediocrity or worse—recharge it with better ways of doing the work at hand.**
>
> —Sutton and Rao (2014, p. 1)

This is how Robert Sutton and Huggy Rao (2014) open their influential book *Scaling Up Excellence*. Scaling is an applied version of the challenge of generalization. Scholars worry about generalizing findings. Philanthropic foundations, policymakers, and social innovators worry about spreading effective programs. Sutton and Rao identify five principles to guide scaling. How did they do it?

Sutton and Rao (2014) focused on two goals:

> Uncovering the most *rigorous* evidence and theory we could find and generating observations and advice that were *relevant* to people who were determined to scale up excellence.

This meant bouncing back and forth between

> the clean, careful, and orderly world of theory and research—that rigor we love so much as academics—and the messy problems, crazy constraints, and daily twists and turns that are relevant to real people as they strive and struggle to spread excellence to those who need it. (p. 298)

Seven Years of Inquiry

Sutton and Rao (2014) report that they began by gathering ideas and evidence, a process that took years.

> We did case studies, reviewed theory and research, and huddled to develop insights about scaling challenges and how to overcome them. Little by little, this process changed from a private conversation between the two of us to ongoing conversations about scaling with an array of smart people. We were at the center of this process: making decisions about which leads, stories, and evidence to pursue; choosing which to keep, discard, or save for later; and weaving them together into (we hope) a coherent form. (p. 299)

Sutton and Rao (2014) then analyzed the evidence to reach preliminary conclusions. As conclusions emerged, they presented what they had found to people who had read their prior publications and/or attended their classes and speeches. They recruited knowledgeable and thoughtful people to review, question, and enhance their work.

> This book is best described as the product of years of give-and-take between us and many thoughtful people, not as an integrated perspective that we constructed in private and are now unveiling for the first time. Hundreds of people played direct roles in helping us, and thousands more played indirect roles—even if they didn't realize it. (p. 299)

To speak to issues of rigor and credibility, Sutton and Rao (2014) have distilled their inquiry process into seven core methods, each of which they elaborate in the methodological appendix of the book.

> 1. Combing through research from the behavioral sciences and beyond
>
> 2. Conducting and gathering detailed case studies
>
> 3. Brief examples from diverse media sources
>
> 4. Targeted interviews as unplanned conversations
>
> 5. Presenting emerging scaling ideas to diverse audiences
>
> 6. Teaching a "*Scaling Up Excellence*" class to Stanford graduate students
>
> 7. Participation in and observation of scaling at the *Stanford* school (an executive professional development program) (pp. 301–306)

What emerges from their description of their inquiry methods is a portrayal of an ongoing, generative, and iterative process of integrating theory, research, and practice around gathering and making sense of the evidence and, ultimately, distilling what

they found into principles. The principles constitute a form of generalized guidance derived from and based on lessons. Remember, earlier I postulated that lessons lead to principles. The book opens with the four lessons they identified that became the basis for formulating their five scaling principles. Here's how Sutton and Rao (2014) describe that connection and the first lesson, which is the basis for treating the principles as generalizations.

> Our first big lesson is that, although the details and daily dramas vary wildly from place to place, the similarities among scaling challenges are more important than the differences. The key choices that leaders face and the principles that help organizations scale up without screwing up are strikingly consistent. (p. xi)

Why Principles?

The seven years of inquiry described by Sutton and Rao (2014) generated five principles. Why principles? Because people engaged in a scaling initiative cannot simply look up some right answers and apply them. There is no recipe.

> In the case of scaling, there are so many different aspects of the challenge, and the right answers vary so much across teams, organizations, and industries (and even across challenges faced by a single team or organization), that it is impossible to develop a useful "paint by numbers" approach. Regardless of how many cases, studies, and books (including this one) you read, success at scaling will always depend on making constantly shifting, complex, and not easily codified judgments. (p. 298)

Principles guide judgment. Context informs judgment. Qualitative inquiry generates principles, and then further qualitative inquiry, in a specific context, illuminates that context so that the principles can be interpreted and applied appropriately within that particular context. That process involves both extrapolation and assessing transferability, the qualitative approach to the challenge of generalizing.

Perspectives on Generalizability: A Review

Four core epistemological issues are at the center of debates about the credibility and utility of qualitative inquiry: (1) judging the quality of findings, (2) inferring causality (the challenge of attribution), (3) the validity of generalizations, and (4) determining what is true.

SIDEBAR

FROM LESSONS TO PRINCIPLES: SCALING THE TRANSFORMATIVE CHANGE INITIATIVE

Started in 2012, the Transformative Change Initiative (TCI) assists community colleges in scaling up innovation: "evidence-based strategies to improve student outcomes and program, organization, and system performance." The TCI evaluation team reviewed case studies of effective innovations, extracted themes and lessons from those separate evaluations of diverse programs, and generated seven principles to guide the next stage of innovation.

> The TCI Framework presents the rationale and guiding principles for scaling innovation in the community college context. It is important to link scaling to guiding principles because principles provide direction rather than prescription. They represent the intentionality of the innovation in ways that often allow for multiple actions (practices) to take place. Principles provide "guidance for action in the face of complexity" so that adaptation can occur in ways that achieve the intended outcome.
>
> The theory of change for TCI suggests scaling happens most successfully when practitioners apply guiding principles to their implementation and scaling efforts. In this view, scaling is not so much about replicating what others assert is good practice, which is a classic theory of scaling, but about practitioners and stakeholders becoming instrumental to the scaling process by igniting a chain of actions, reactions, and outcomes that reflect and ultimately reshape the context. To make this happen, practitioners need to
>
> - be aware of the principles that guide the changes they are making to their practice,
> - reflect those principles in implementation over time, and
> - measure and assess whether the changes are producing the intended improved performance.
>
> —Bragg et al. (2014, p. 6)
> *Transformative Change Initiative*

The first part of this chapter dealt with the issue of quality by examining alternative criteria for judging quality (Modules 76 and 77). Chapter 8 included an extensive discussion of causal inference (pp. 582–595). This module has been examining perspectives on making generalizations. The next and final module will take up the issue of determining what is true. This

EXHIBIT 9.14 Twelve Perspectives on and Approaches to Generalization of Qualitative Findings

INQUIRY PERSPECTIVE	APPROACH TO GENERALIZATION	ELABORATION
1. Traditional scientific research approaches: • Grounded theory • Analytic induction • Qualitative comparative analysis	Generalizations must be theory based: rigorous and systematic comparisons of observed patterns with theoretical propositions.	Qualitative inquiry can contribute generalizable knowledge by generating, testing, and validating theory.
2. Realism	Generalizations depend on purposeful theoretical sampling.	"By linking decisions about whom or what to sample both empirical and theoretical considerations are combined and claims can be made about how the chosen sample relates to a wider universe or population" (Emmel, 2013, p. 60).
3. Constructivism	*Transferability* of findings from particular cases to others based on similarity of context and conditions—also called *inferential generalization* (Lewis & Ritchie, 2003)	Eschew "generalization" in favor of assessing transferability based on in-depth knowledge about the cases studied that provides a basis for assessing the relevance of findings to other similar cases (Lincoln & Guba, 1985).
4. In-depth case study particularity	Focus first on in-depth particularity. Do justice to the case. The issue here is *internal generalization*: "generalizing within the setting . . . studied to people, events, and settings that were not directly observed or interviewed . . . ; the extent to which the times and places observed may differ from those that were not observed, either because of sampling or because of the observation itself" (Maxwell, 2012, p. 142).	"Particularization does deserve praise. . . . What becomes useful understanding is a full and thorough knowledge of the particular, recognizing it also in new and foreign contexts. That knowledge is a form of generalization too . . . , arrived at by recognizing the similarities of objects and issues in and out of context and by sensing the natural covariations of happenings. To generalize this way is to be both intuitive and empirical" (Stake, 1978, p. 6).
5. Social construction	Reflective, experiential, and socially shared generalizations: People naturally make comparisons, which become a "natural" form of generalizing among a group of people. Qualitative cases can enhance those comparisons through the depth and detail that enhances understanding.	*Naturalistic generalizations.* The kind of learning that ordinary people take from their encounters with specific case studies: "The 'vicarious experience' that comes from reading a rich case account can contribute to the social construction of knowledge which, in a cumulative sense, builds general, if not necessarily generalizable knowledge" (Stake, 1995, p. 38).

(Continued)

(Continued)

INQUIRY PERSPECTIVE	APPROACH TO GENERALIZATION	ELABORATION
6. Phenomenology	*Essence.* "A unified statement of the essences of the experience of the phenomenon as a whole. . . . Essence . . . means that which is common or universal, the condition or quality without which a thing would not be what it is" (Moustakas, 1994, p. 100).	Essence emerges from a synthesis of meanings, a reduction of variation to what is essential. Essence integrates and supersedes individual experiences. "The essences of any experience are never totally exhausted. The fundamental textural-structural synthesis represents the essences at a particular time and place from the vantage point of an individual researcher following an exhaustive imaginative and reflective study of the phenomenon" (Moustakas, 1994, p. 100).
7. Ethnography	*Conceptual generalization*: describing both the common and variable meanings and manifestations of universal concepts across cultures, e.g., kinship, conflict, religion, coming of age, etc.	Connecting the microscopic or situation-specific findings of a particular ethnography to more general understandings of culture (Geertz, 1988, 2001) constitutes a form of ethnographic generalization. It involves seeking "the pattern that connects" (Bateson, 1977, 1988).
8. Pragmatism	*Extrapolations.* Modest, practical speculations on the likely applicability of findings to future times and other situations under similar, but not identical, conditions; allows for more interpretive flexibility than a direct transferability assessment; as interested in temporal application of findings (applying what was learned to the future) as in applications in other places.	Extrapolations are logical, thoughtful, case derived, problem oriented, and future oriented rather than statistical and probabilistic: what of practical value has been learned that can be extrapolated to guide future actions, whether in the same place or a different one. Extrapolations can be particularly useful when based on information-rich samples targeted to specific concerns (Cronbach, 1980).
9. Program evaluation and policy analysis	*Lessons.* Qualitative evaluation's contribution to general knowledge takes the form of lessons identified in one evaluation (or a cluster of evaluations) that are offered for application to other places and future programs.	Patterns of effectiveness identified from cross-case analysis of different programs are analyzed to extract common lessons. A classic example is Schorr's (1988) "Lessons of Successful Programs," serving high-risk children and families in poverty. Here is a policy example: *Learning From Iraq* (Special Inspector General for Iraq Reconstruction, 2013).
10. Systems and complexity	*Principles.* Dynamic and complex systems defy simple empirical generalizations. Instead, principles can be identified that inform future systems analyses and guide innovation in complex situations.	Case-based principles provide guidance for adaptive action in the face of complexity. Adaptive action through principles contrasts high-fidelity replication of standardized models. Principles emphasize contextual sensitivity and situational analysis (Eoyang & Holladay, 2013; Patton, 2011).

INQUIRY PERSPECTIVE	APPROACH TO GENERALIZATION	ELABORATION
11. Artistic, evocative representations	Emotional connections and empathy are a form of human generalizability based on shared feelings. *Interpretive interactionism* (Denzin, 1989b) involves intersubjective understandings and feelings.	Finding shared meaning in stories and artistic works (as representations of qualitative findings) moves people from their isolated, particular experience to a more general experience and understanding of the human condition. Emotional resonance among humans (Denzin, 2009) is a form of empathic generalization.
12. Postmodernism	All knowledge is local, specific, and immediate. Generalizations, either empirical or theoretical, are impossible and undesirable.	"Postmodernism is characterized by its distrust of and incredulity toward all 'totalizing' discourses or metanarratives" (Schwandt, 2007, p. 235). This discounts the validity of generalizations, theories, and predictions of any kind, "alerting us to postmodernism's nihilistic tendencies" (Gubrium & Holstein, 2003, p. 5).

SIDEBAR

"ALL GENERALIZATIONS ARE FALSE."

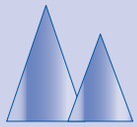
"All generalizations are false, including this one"—doesn't clarify much of anything.

"All generalizations are false, including this one" leads logically to "Some generalizations are true." If you wish to trace this error back, consider "All Cretans lie," uttered by a Cretan. It can't be true, but it can be false. What's interesting is that it not only leads to "Some Cretans tell the truth," but it also leads to the conclusion that the Cretan speaking is not one of them.

—Errol Morris (2014)
Documentary filmmaker and philosopher

MODULE

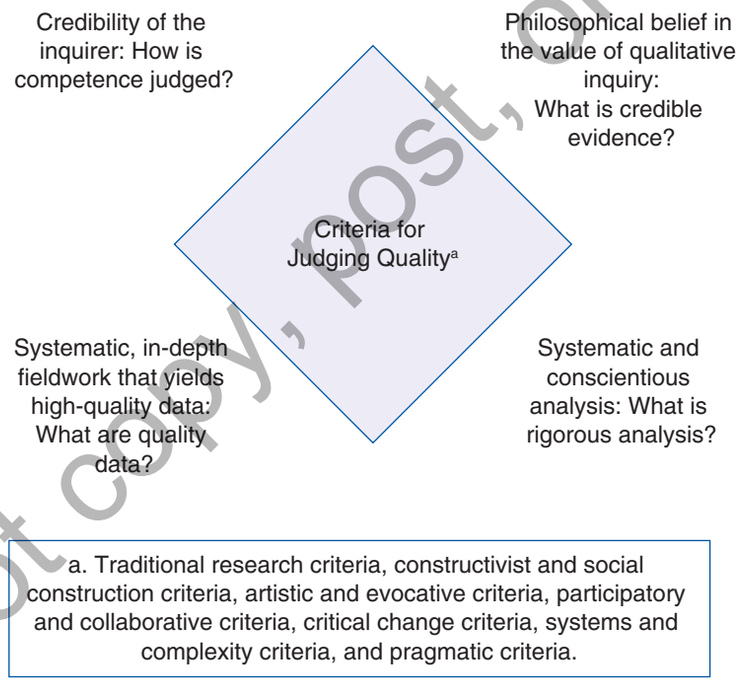
82 Enhancing the Credibility and Utility of Qualitative Inquiry by Addressing Philosophy of Science Issues



module concludes with a summary of perspectives on and approaches to generalization in Exhibit 9.14.

We come now to the fourth and final dimension of credibility. Let's review. The first dimension is *systematic, in-depth fieldwork that yields high-quality data.* The second dimension that informs judgments of credibility is *systematic and conscientious analysis*. The third concerns judgments about the *credibility of the researcher*, which depends on training, experience, track record, status, and presentation of self. Now, to conclude, we take up the issue of *philosophical belief in the value of qualitative inquiry*, that is, a fundamental appreciation of naturalistic inquiry, qualitative methods, inductive analysis, purposeful sampling, and holistic thinking. Exhibit 9.15 graphically depicts these four dimensions of credibility. In the center of the graphic are the alternative criteria for judging quality that opened this chapter: traditional scientific research criteria, constructivist and social construction criteria, artistic and evocative criteria, participatory

EXHIBIT 9.15 Criteria for Judging Quality



a. Traditional research criteria, constructivist and social construction criteria, artistic and evocative criteria, participatory and collaborative criteria, critical change criteria, systems and complexity criteria, and pragmatic criteria.

and collaborative criteria, critical change criteria, systems and complexity criteria, and pragmatic criteria.

Philosophical belief in the value of qualitative inquiry is a prime determinant of credibility—and a matter of debate and controversy. Given the often-controversial nature of qualitative findings and the necessity, on occasion, to be able to explain and even defend the value and appropriateness of qualitative inquiry, this module will briefly discuss some of the most contentious issues. The selection of which philosophy of science issues to address in this closing section of the book is based on the workshops I regularly teach on qualitative evaluation methods. In those two- and three-day courses, which typically include participants from around the world, I reserve the final

afternoon for open-ended exchanges about whatever matters of interest and concern participants want to raise. By then, we have covered types and applications of qualitative inquiry, design options, purposeful sampling approaches, fieldwork techniques, observational methods, interviewing skills, how to do systematic and rigorous analysis, and ethical standards. Inevitably, questions come pouring forth about the paradigms debate, political considerations, and fundamental doubts participants encounter about the legitimacy of qualitative inquiry. I'll reproduce the questions that arise and offer my responses.

Paradigms question: *Why are qualitative methods so controversial? I just want to interview people, see what they say, analyze the patterns, and report my findings? I don't want to debate paradigms. Do we really have to deal with paradigms stuff?*

You have to deal with what constitutes credible evidence. What constitutes credible evidence is a matter of debate among both scientists and nonscientists. While not always framed as a paradigms debate, and there are disagreements about what a paradigm is and whether it's a useful concept, I think framing the controversy as a paradigms debate is both accurate and illuminating. In Chapter 3, I discussed the qualitative/quantitative paradigms debate at some length (see pp. 87–95), including an MQP Rumination against designating randomized controlled trials as the "gold standard." In this module, I'm going to focus specifically on how that debate affects credibility and utility.

Paradigms are a way of distinguishing different perspectives in science about how best to study and understand the world. The debate sometimes takes the form of natural science versus social science, qualitative versus quantitative methods, behavioral psychology versus phenomenology, positivism versus constructivism, or realism versus interpretivism. How the debate is framed depends on the perspectives that people bring to it and the language available to them to talk about it. Whatever the terminology and labels for contrasting points of view, the debate is rooted in philosophical differences about the nature of reality and epistemological differences in what constitutes knowledge and how it is created. The *paradigms debate,* whatever form it takes, affects credibility and utility when particular worldviews are pitted against one another at the intersection of philosophy and methods to determine what kinds of evidence are acceptable, believable, and useful.

You may be able to carry out a qualitative study without ever addressing the issue of paradigms. But you ought to know enough about the debate and its implications, it seems to me, to address the issue if it comes up. I would alert those new to the debate that it has been and can be intense, divisive, emotional, and rancorous. And to those experienced in and tired of the debate, let me say that I've followed it, and been personally engaged in it, for more than 40 years. I've watched the debate ebb and flow, take on new forms, and attract new advocates and adversaries. But it doesn't go away. The paradigms debate is an epistemological phoenix that emerges anew when fires of dissent mellow into become dying embers only to flame again on new winds of contention. I doubt that you can use qualitative methods without encountering and needing to deal with some aspects of the debate. As I have illustrated throughout this chapter, both scientists and nonscientists hold strong opinions about what constitutes credible evidence. Those opinions are paradigm derived and paradigm dependent because a paradigm constitutes a worldview built on epistemological assumptions, preferred definitions of key concepts, comfortable habits, entrenched values defended as truths, and beliefs offered up as evidence. As such, paradigms are deeply embedded in the socialization of adherents and practitioners, telling them what is important, legitimate, and reasonable.

So be prepared to address controversies and competing perspectives about what constitutes credible evidence even if it doesn't come cloaked in the guise of a paradigms debate. Moreover, these are not simply matters of academic debate. They have entered the public policy arena as matters of political debate.

Politics of evidence question**:** *What makes research methods a matter of concern for politics and politicians?*

In the public policy arena, advocates of randomized control trials are organized and funded to lobby the U.S. Congress to put their paradigm preferences into legislation (Coalition for Evidence-Based Policy, 2014). They have communications experts who supply reporters with positive news accounts (e.g., Keating, 2014; Kolata, 2013, 2014). On the other side, there are strong political advocacy statements for alternative paradigms: *The Qualitative Manifesto* (Denzin, 2010), *Qualitative Inquiry and the Conservative Challenge* (Denzin & Giardina, 2006), and *Qualitative Inquiry and the Politics of Evidence* (Denzin & Giardina, 2008). Ray Pawson (2013) has produced *A Realist Manifesto*. But there is no organized and funded lobbying effort on behalf of qualitative, mixed-methods, and/or realist approaches. So guess which group is successful in getting its paradigm legitimated and funded in legislation? *Hint:* It's not the qualitative manifesto.

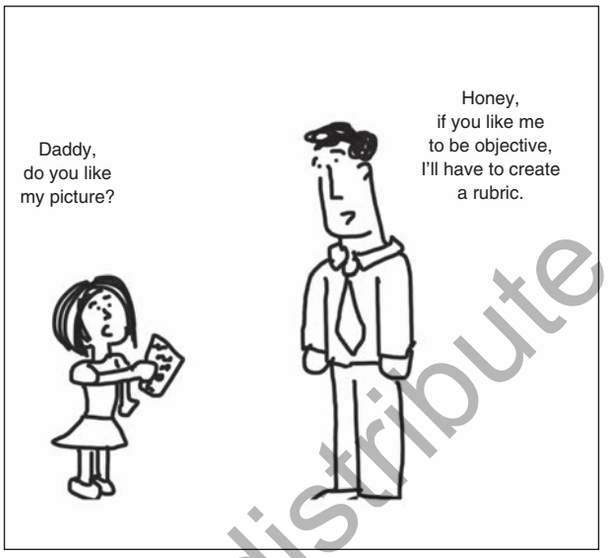
Objectivity question**:** *Doesn't the paradigms debate come down to objectivity versus subjectivity?*

French philosopher Jean-Paul Sartre once observed that "words are loaded pistols." The words "objectivity" and "subjectivity" are bullets people arguing fire

SIDEBAR

DIFFERENT MEANINGS AND USES OF OBJECTIVITY

1. *Objective person.* Unbiased, open-minded, and neutral
2. *Objective process.* Follow, document, and report procedures that do not predetermine results
3. *Objective statement.* Just the facts, unvarnished, put forward by an objective person following an objective process
4. *Objective reality.* Belief that there is knowable, absolute reality
5. *Objective scientific claim.* Findings subjected to scientific peer review by members of a discipline capable of judging the extent to which a claim has been produced by appropriate scientific methods and analysis
6. *Objective methods.* A design, data collection procedures, and analysis that follow accepted inquiry norms of a scientific discipline
7. *Objective measure.* The extent to which a given number can be interpreted as indicating the same amount of the thing measured, across persons or thing measured, using a validated and reliable instrument
8. *Objective decisions.* Fair and balanced judgment based on preponderance of evidence presented and explicit; transparent criteria for weighing the evidence



SOURCE: © Chris Lysy—freshspectrum.com

at each other. It's true that objectivity is held in high esteem. Science aspires to objectivity and a primary reason why decision makers commission an evaluation is to get objective data from an independent source external to the program being evaluated. The charge that qualitative methods are inevitably "subjective" casts an aspersion connoting the very antithesis of scientific inquiry. Objectivity is traditionally considered the sine qua non of the scientific method. To be subjective means to be biased, unreliable, and irrational. Subjective data imply opinion rather than fact, intuition rather than logic, impression rather than confirmation. Chapter 2 briefly discussed concerns about objectivity versus subjectivity, but I return to the issue here to address how these concerns affect the credibility and utility of qualitative analysis.

Let's take a closer look at the objective/subjective distinction. The conventional means for controlling subjectivity and maintaining objectivity are the methods of quantitative social science: distance from the setting and people being studied, standardized quantitative measures, formal operational procedures, manipulation of isolated variables, and randomized controlled experimental designs. Yet the ways in which measures are constructed in psychological tests, questionnaires, cost–benefit indicators, and routine management information systems are no less open to the intrusion of biases than making observations in the field or asking questions in interviews. Numbers do not protect against bias; they merely disguise it. All statistical data are based on *someone's* definition of what to measure and how to measure it. An "objective" statistic like the consumer price index is really made up of very subjective decisions about what consumer items to include in the index. Periodically, government economists change the basis and definition of such indices.

Philosopher of science Michael Scriven (1972a) has insisted that quantitative methods are no more synonymous with objectivity than qualitative methods are synonymous with subjectivity:

> Errors like this are too simple to be explicit. They are inferred confusions in the ideological foundations of research, its interpretations, its application. . . . It is increasingly clear that the influence of ideology on methodology and of the latter on the training and behavior of researchers and on the identification and disbursement of support is staggeringly powerful. Ideology is to research what Marx suggested the economic factor was to politics and what Freud took sex to be for psychology. (p. 94)

Scriven's (1972a) lengthy discussion of objectivity and subjectivity in educational research deserves careful reading by students and others concerned by

this distinction. He skillfully detaches the notions of objectivity and subjectivity from their traditionally narrow associations with quantitative and qualitative methodology, respectively. He presents a clear explanation of how objectivity has been confused with consensual validation of something by multiple observers. Yet a little research will yield many instances of "scientific blunders" (Dyson, 2014; Livio, 2013; Youngson, 1998) where the majority of scientists were factually wrong while *one* dissenting observer described things as they really were (Kuhn, 1970).

Qualitative rigor has to do with the quality of the observations made by an inquirer. Scriven (1972a) emphasizes the importance of being factual about observations rather than being distant from the phenomenon being studied. *Distance does not guarantee objectivity; it merely guarantees distance*. Nevertheless, in the end, Scriven (1998) still finds the ideal of objectivity worth striving for as a counter to bias, and he continues to find the language of objectivity serviceable.

In contrast, Lincoln and Guba (1986), as noted earlier, have suggested replacing the traditional mandate to be objective with an emphasis on *trustworthiness* and *authenticity* by being balanced, fair, and conscientious in taking account of multiple perspectives, multiple interests, multiple experiences, and diverse constructions of realities. Guba (1981) suggested that researchers and evaluators can learn something about these attributes from the stance of investigative journalists.

> Journalism in general and investigative journalism in particular are moving away from the criterion of objectivity to an emergent criterion usually labeled "fairness" . . . Objectivity assumes a single reality to which the story or evaluation must be isomorphic; it is in this sense a one-perspective criterion. It assumes that an agent can deal with an objective (or another person) in a nonreactive and noninteractive way. It is an absolute criterion.
>
> Journalists are coming to feel that objectivity in that sense is unattainable. . . .
>
> Enter "fairness" as a substitute criterion. In contrast to objectivity, fairness has these features:
>
> - It assumes multiple realities or truths—hence a test of fairness is whether or not "both" sides of the case are presented, and there may even be multiple sides.
>
> - It is adversarial rather than one-perspective in nature. Rather than trying to hew the line with the truth, as the objective reporter does, the fair reporter seeks to present each side of the case in the manner of an advocate—as, for example, attorneys do in making a case in court. The presumption is that the public, like a jury, is more likely to reach an equitable decision after having heard each side presented with as much vigor and commitment as possible.
>
> - It is assumed that the subject's reaction to the reporter and interactions between them heavily determines what the reporter perceives. Hence one test of fairness is the length to which the reporter will go to test his own biases and rule them out.
>
> - It is a relative criterion that is measured by *balance* rather than by isomorphism to enduring truth. (pp. 76–77)

But times change, and Guba would be unlikely to use the language of "fairness and balance" now that the most politically conservative and deliberately biased American television channel has adopted that phrase as its brand. *Fairness and balance* has become a euphemism for *prejudiced and one-sided.* Objectivity has also taken on unfortunate political and cultural connotations in some quarters, meaning uncaring, unfeeling, disengaged, and aloof. What about subjectivity, the constructivist badge of honor?

Subjectivity Deconstructed

In public discourse, it is not particularly helpful to know that philosophers of science now typically doubt the possibility of anyone or any method being totally "objective." But subjectivity fares even worse. Even if acknowledged as inevitable (Peshkin, 1988), or valuable as a tool to understanding (Soldz & Andersen,

2012), subjectivity carries such negative connotations at such a deep level and for so many people that the very term can be an impediment to mutual understanding. For this and other reasons, as a way of elaborating with any insight the nature of the research process, the notion of subjectivity may have become as useless as the notion of objectivity.

> The death of the notion that objective truth is attainable in projects of social inquiry has been generally recognized and widely accepted by scholars who spend time thinking about such matters. . . . I will take this recognition as a starting point in calling attention to a second corpse in our midst, an entity to which many refer as if it were still alive. Instead of exploring the meaning of subjectivity in qualitative educational research, I want to advance the notion that following the failure of the objectivists to maintain the viability of their epistemology, the concept of subjectivity has been likewise drained of its usefulness and therefore no longer has any meaning. Subjectivity, I feel obliged to report, is also dead. (Barone, 2000, p. 161)

But for other qualitative researchers, subjectivity is not so much about philosophy of science as it is about using one's own experience to make sense of the world through reflexivity (Connolly & Reilly, 2007). That perspective, once entertained, can lead from a focus on the researcher's subjectivity as a window into sense making to ***shared meaning making***: intersubjectivity.

Intersubjectivity

> "Subjective" versus "objective" no longer makes sense, since everyone involved is a subject. . . . Human Social Research is *intersubjective* . . . built from encounters among subjects, including researchers who, like it or not, are also subjects. (Agar, 2013, pp. 108–109)

Eschewing both objectivity and subjectivity, intersubjectivity focuses on knowledge as socially constructed in human interactions. Human science research, what anthropologist Michael Agar (2013) calls the *Lively Science*, requires "human social relationships in order to happen at all. They are *intersubjective* sciences. They require social relationships with those who support the science, those who do it, those who serve as subjects of it, and those who consume it" (p. 215).

> The difficult judgment call for the researcher is this: To some extent he or she *should* translate his or her own framework and jointly build a framework for communication with subjects of all those different types. . . . The bedrock of intersubjective research isn't to preach or to lecture, but rather to learn and to communicate the results, though not at the price of abandoning the core principles of the science. The pressure always exists to achieve a balance, and a researcher always has to make the call of how much and in what way to handle it.
>
> *This fact has to be part of the* science, *not to mention a central part of training for human social researchers*. How to navigate this ambiguous territory with professional integrity and product quality is a neglected topic, a neglect understandable in light of academic traditions where one could assume that whatever the dissertation committee or disciplinary peers would like was the right thing to do. That isolation is no longer possible. In my view, taking human social research out into the world makes it more difficult, more interesting, more intellectually challenging, and of higher moral value than it has ever been. (pp. 215–216)

Empathic Neutrality

No consensus about substitute terminology has emerged. I prefer empathic neutrality, one of the 12 qualitative themes that I presented in Chapter 2.

> While empathy describes a stance toward the people we encounter in fieldwork, calling on us to communicate interest, caring, and understanding, neutrality suggests a stance toward their thoughts, emotions, and behaviors, a stance of being nonjudgmental. Neutrality can actually facilitate rapport and help build a relationship that supports empathy by disciplining the researcher to be open to the other person and nonjudgmental in that openness.

(See pp. 57–62 for the full discussion of empathic neutrality.)

Open-Mindedness and Impartiality

I have evaluation colleagues who simply describe themselves as open-minded, which seems to satisfy most lay people. The political nature of evaluation means that individual evaluators must make their own peace with how they are going to describe what they do. The meaning and connotations of words like *objectivity*, *subjectivity*, *neutrality*, and *impartiality* will have to be worked out with particular stakeholders in specific evaluation settings. In her leadership role in evaluation in the U.S. federal government, former AEA president Eleanor Chelimsky emphasized her unit's independence and impartiality. The perception of impartiality, she has explained, is at least as important as methodological rigor in highly political environments. Credibility, and therefore utility, are affected by "the steps

we take to make and explain our evaluative decisions, [and] also intellectually, in the effort we put forth to look at all sides and all stakeholders of an evaluation" (Chelimsky, 1995, p. 219; see also Chelimsky, 2006).

I think it is worth noting that the official Program Evaluation Standards (Joint Committee on Standards, 2010) do not call for objectivity. The standards have been guiding evaluation practice for nearly four decades. They were originally formulated by social scientists and evaluators representing all the major disciplinary associations. They have twice gone through major review processes. The language used, therefore, has been thoroughly vetted. The standards call for evaluations to be credible, systematic, accurate, useful, accurate, and dependable, but not objective. The term *objectivity* has become a lightning rod attracting epistemological paradigms debate and therefore not useful as a standard for evaluation in the American context. In contrast, the international Quality Standards for Development Evaluation define evaluation as "objective assessment" (OECD-DAC, 2010, p. 5). Different context, different language.

Given the seven different sets of criteria for judging the quality of qualitative inquiry I identified at the beginning of this chapter, and the terms associated with each, it seems unlikely that a consensus about terminology is on the horizon. The methodological and scientific Tower of Babel stands tall and casts a long shadow. But the different perspectives on and uses of terms can be liberating because they opens up the possibility of getting beyond the meaningless abstractions and heavy-laden connotations of objectivity and subjectivity to move instead toward carefully selecting *descriptive* methodological language that best describes your own inquiry processes and procedures. That is, don't label those processes as "objective," "subjective," "intersubjective," "trustworthy," or "authentic." Instead, eschew overarching labels. Describe how you approach your inquiry, what you bring to your work, and how you've reflected on what you do, and then let the reader be persuaded, or not, by the intellectual and methodological rigor, meaningfulness, value, and utility of the result. In the meantime, be very careful how you use particular terms in specific contexts. Words are bullets. They are also landmines. I end this diatribe with a cautionary tale about being sensitive to the cultural context within which terms are used.

During a tour of America, former British prime minister Winston Churchill attended a buffet luncheon at which chicken was served. As he returned to the buffet for a second helping he asked, "May I have some more breast?"

His hostess, looking embarrassed, explained that "in this country we ask for white meat or dark meat."

Churchill, taking the white meat he was offered, apologized and returned to his table.

The next morning the hostess received a beautiful orchid from Churchill with the following card: "I would be most obliged if you would wear this on your white meat."

SIDEBAR

A REALIST PERSPECTIVE ON OBJECTIVITY

> Evaluation cannot hope for perfect objectivity but neither does this mean it should slump into rampant subjectivity. We cannot hope for absolute cleanliness but this does not require us to enjoy a daily roll in the manure. The alternative to these two termini is for evaluation to embrace the goal of being "validity increasing" . . .
>
> Skepticism . . . , in its English spelling, . . . constitutes the final desideratum of evaluation science.
>
> Organised scepticism means that any scientific claim must be exposed to critical scrutiny before it becomes accepted. . . . What counts is the depth of critical scrutiny applied to the inferences drawn from any inquiry. And this level of attention depends, in turn, on the presence of a collegiate group of stakeholders and their willingness to put each other's work under the microscope.
>
> —Ray Pawson (2013, p. 107)
> *The Science of Evaluation: A Realist Manifesto*

Truth and reality question: *"I don't understand this talk about multiple realities and different truths for different people. If research is anything, it ought to be about getting at true reality. I know you like quotes, so here's one of my favorite quotes for you, from George Orwell: 'In a time of universal deceit—telling the truth is a revolutionary act. 'I think we ought to be research revolutionaries and speak the truth. In fact, the mantra of evaluation is: Speak truth to power. So, truth or not truth?"*

It's an important question. Certainly, there are a lot of quotes about truth. This is a thick book, and it could contain nothing but quotes about truth, which would serve to illustrate its evasiveness. Let me offer a quote from the great comedian Lily Tomlin, who, playing the character of a little girl accused by a scolding adult of making things up, responded thus:

> Lady, I do not make up things. That is lies. Lies are not true. But the truth could be made up if you know how. And that's the truth.

Or consider this observation by Thomas Schwandt, a philosopher of science and professional evaluator, who has spent much of a distinguished career grappling with this very issue. His conclusion:

> **TRUTH** is one of the most difficult of all philosophical topics, and controversies surrounding the nature of truth lie at the heart of both apologies for and criticisms of varieties of qualitative work. Moreover, truth is intimately related to questions of *meaning*, and establishing the nature of that relationship is also complicated and contested.
>
> There is general agreement that *what* is true or what carries truth are statements, propositions, beliefs, and assertions, but *how* the truth of same is established is widely debated. (Schwandt, 2007, p. 300)

Schwandt presents 10 different philosophical orientations to and theories about truth: (1) correspondence, (2) consensus, (3) coherence, (4) contextualist, (5) pragmatic, (6) hermeneutic, (7) critical theory (Foucault), (8) realist, (9) constructivist, and (10) objectivist theory. Pick your poison—or truth. We won't resolve the debate here. Not even close. Nor will others for, to add yet another quote to the collection, here's cynic Ambrose Bierce's (1999) assessment:

> Discovery of truth is the sole purpose of philosophy, which is the most ancient occupation of the human mind and has a fair prospect of exiting with increasing activity to the end of time. (p. 201)

Since we can't resolve the nature of truth, indulge me in a story that illustrates why it may be important to have figured out where you, yourself, stand on matters of truth. Following a presentation of evaluation findings at a public school board meeting, I was asked by the school district's internal evaluator, "Do you, as a qualitative researcher, swear to tell the truth, the whole truth and nothing but the truth?" The question was meant to embarrass me. The researcher had an article I had written attacking overreliance on standardized tests for school evaluations and another advocating soliciting multiple perspectives from parents, teachers, students, and community members about their experiences with the school district to document diverse perspectives. In that article, and earlier editions of this book, I had expressed doubt about the utility of truth as a criterion of quality and I suspected that he hoped to lure me into an academic-sounding, arrogant, and philosophical discourse on the question "What is truth?" in the expectation that the public officials present would be alienated and dismiss my presentation. So when he asked, "Do you, as a qualitative researcher, swear to tell the truth, the whole truth and nothing but the truth?" I did *not* reply, "That depends on what *truth* means." I said simply, "Certainly I promise to respond honestly." Notice the shift from truth to honesty.

The researcher applying traditional social science criteria might respond, "I can show you truth insofar as it is revealed by the data."

The constructivist might answer, "I can show you multiple truths."

The artistically inclined might suggest that "beauty is truth." And "fiction often reveals truth better than nonfiction."

The critical theorist could explain that "truth depends on one's consciousness."

The participatory qualitative inquirer would say, "We create truth together."

The critical change activist might say, "I offer you praxis. Here is where I take my stand. This is true for me."

The pragmatic evaluator might reply, "I can show you what is useful. What is useful is true."

Indeed, in this vein, Exhibit 9.7, in presenting the seven sets of criteria for judging quality, offers a political campaign button about TRUTH for each (pp. 680–681).

By the way, I noted earlier that the Program Evaluation Standards do not use the language of objectivity, but the "the Accuracy Standards are intended to increase the dependability and *truthfulness* [italics added] of evaluation representations" (Joint Committee on Standards, 2010). *Note: Truthfulness* is not TRUTH. You could do a little hermeneutic work on that distinction, should you be so inclined.

Ironically, it is sometimes easier to determine what is false than what is true. For insights into how the academic peer review process has been distorted and corrupted to generate invalid and untrustworthy results, see the widely cited and influential analysis by Professor of Health Research and Policy at Stanford School of Medicine, John P. A. Loannidis (2005) "Why Most Published Research Findings Are False."

Truth Tests and Utility Tests

Previously I have cited the influential research by Weiss and Bucuvalis (1980) that decision makers apply both "truth" tests and "utility" tests to evaluation. "Truth," in this case, however, means reasonably accurate and credible data (the focus of the program

evaluation standards) rather than data that are true in some absolute sense. Savvy policymakers know better than most the context and perspective-laden nature of competing truths. Qualitative inquiry can present accurate data on various perspectives, including the evaluator's perspective, without the burden of determining that only one perspective must be true.

Evaluation theorist and methodologist Nick Smith (1978), pondering these questions, has noted that to act in the world we often accept either approximations to truth or even untruths.

> For example, when one drives from city to city, one acts as if the earth is flat and does not try to calculate the earth's curvature in planning the trip, even though acting as if the earth is flat means acting on an untruth. Therefore, in our study of evaluation methodology, two criteria replace exact truth as paramount: practical utility and level of certainty. The level of certainty required to make an adequate judgment under the law differs depending on whether one is considering an administrative hearing, an inquest, or a criminal case. Although it seems obvious that much greater certainty about the nature of things is required when legislators set national and educational policy than when a district superintendent decides whether to continue a local program, the rhetoric in evaluation implies that the same high level of certainty is required of both cases. If we were to first determine the level of certainty desired in a specific case, we could then more easily choose appropriate methods. Naturalistic descriptions give us greater certainty in our understanding of the nature of an educational process than randomized, controlled experiments do, but less certainty in our knowledge of the strength of a particular effect. . . . Our first concern should be the practical utility of our knowledge, not its ultimate truthfulness. (p. 17)

In studying evaluation use (Patton, 2008), I found that decision makers did not expect evaluation reports to produce "TRUTH" in any fundamental sense. Rather, they viewed evaluation findings as additional information that they could and did combine with other information (political, experiential, other research, colleague opinions, etc.), all of which fed into a slow, evolutionary process of incremental decision making. Kvale (1987) echoed this interactive and contextual approach to truth in emphasizing the "pragmatic validation" of findings in which the results of qualitative analysis are judged by their relevance to and use by those to whom findings are presented.

This *criterion of utility* can be applied not only to evaluation but also to qualitative analyses of all kinds, including textual analysis. Barone (2000), having rejected objectivity and subjectivity as meaningless criteria in the postmodern age, makes the case for pragmatic utility:

> If all discourse is culturally contextual, how do we decide which deserves our attention and respect? The pragmatists offer the criterion of usefulness for this purpose. . . . An idea, like a tool, has no intrinsic value and is "true" only in its capacity to perform a desired service for its handler within a given situation. When the criterion of usefulness is applied to context-bound, historically situated transactions between itself and a text, it helps us to judge which textual experiences are to be valued. . . . The gates are opened for textual encounters, in any inquiry genre or tradition, that serve to fulfill an important human purpose. (pp. 169–170)

Focusing on the connection between truth tests and utility tests shifts attention back to credibility and quality, not as absolute generalizable judgments but as contextually dependent on the needs and interests of those receiving our analysis. This obliges researchers and evaluators to consider carefully how they present their work to others, with attention to the purpose to be fulfilled. That presentation should include reflections on how your perspective affected the questions you pursued in fieldwork, careful documentation of all procedures used so that others can review your methods for bias, and being open in describing the limitations of the perspective presented. Exhibit 9.16, at the end of this chapter (pp. 736–741), offers an in-depth description of how one qualitative inquirer dealt with these issues in a long-term participant–observer relationship. The exhibit, titled A Documenter's Perspective, is based on her research journal and field notes. It moves the discussion from abstract philosophizing to day-to-day, in-the-trenches fieldwork encounters aimed at sorting out what is true (small *t*) and useful.

Finding TRUTH can be a heavy burden. I once had a student who was virtually paralyzed in writing an evaluation report because he wasn't sure if the patterns he thought he had uncovered were really true. I suggested that he not try to convince himself or others that his findings were true in any absolute sense but, rather, that he had done the best job he could in describing the patterns that appeared to him to be present in the data and that he present those patterns as *his* perspective based on his analysis and interpretation of the data he had collected. Even if he believed that what he eventually produced was Truth, any sophisticated person reading the report would know that what he presented was no more than his perspective, and they would judge that perspective by their own commonsense understandings and use the

SIDEBAR

TRUTH VERSUS RELATIVISM

Postmodern work is often accused of being relativistic (and evil) since it does not advocate a universal, independent standard of truth. In fact, relativism is only an issue for those who believe there is a foundation, a structure against which other positions can be objectively judged. In effect, this position implies that there is no alternative between objectivism and relativism. Postmodernists dispute the assumptions that produce the objectivism/relativism binary since they think of truth as multiple, historical, contextual, contingent, political, and bound up in power relations. Refusing the binary does not lead to the abandonment of truth, however, as Foucault emphasizes when he says, "I believe too much in truth not to suppose that there are different truths and different ways of speaking the truth."

Furthermore, postmodernism does not imply that one does not discriminate among multiple truths, that "anything goes." . . . If there is no absolute truth to which every instance can be compared for its truth-value, if truth is instead multiple and contextual, then the call for ethical practice shifts from grand, sweeping statements about truth and justice to engagements with specific, complex problems that do not have generalizable solutions. This different state of affairs is not irresponsible, irrational, or nihilistic. . . . As with truth, postmodern critiques argue for multiple and historically specific forms of reason. (St. Pierre 2000, p. 25)

information according to how it contributed to their own needs.

As one additional source of reflection on these issues, perhaps the following Sufi story will provide some guidance about the difference between truth and perspective. Sagely, in this encounter, Nasrudin gathers data to support his proposition about the nature of truth. Here's the story.

Mulla Nasrudin was on trial for his life. He was accused of no less a crime than treason by the king's ministers, wise men charged with advising on matters of great import. Nasrudin was charged with going from village to village inciting the people by saying, "The king's wise men do not speak truth. They do not even know what truth is. They are confused." Nasrudin was brought before the king and the court. "How do you plead, guilty or not guilty?"

"I am both guilty and not guilty," replied Nasrudin.

"What, then, is your defense?"

Nasrudin turned and pointed to the nine wise men who were assembled in the court. "Have each sage write an answer to the following question: 'What is water?'"

The king commanded the sages to do as they were asked. The answers were handed to the king, who read to the court what each sage had written.

The first wrote, "Water is to remove thirst."

The second, "It is the essence of life."

The third, "Rain."

The fourth, "A clear, liquid substance."

The fifth, "A compound of hydrogen and oxygen."

The sixth, "Water was given to us by God to use in cleansing and purifying ourselves before prayer."

The seventh, "It is many different things—rivers, wells, ice, lakes, so it depends."

The eighth, "A marvelous mystery that defies definition."

The ninth, "The poor man's wine."

Nasrudin turned to the court and the king: "I am guilty of saying that the wise men are confused. I am not, however, guilty of treason because, as you see, the wise men are confused. How can they know if I have committed treason if they cannot even decide what water is? If the sages cannot agree on the truth about water, something which they consume every day, how can one expect that they can know the truth about other things?"

The king ordered that Nasrudin be set free.

SIDEBAR

TRUE FACTS VERSUS TRUE THEORIES

Facts and theories are born in different ways and are judged by different standards. Facts are supposed to be true or false. They are discovered by observers or experimenters. A scientist who claims to have discovered a fact that turns out to be wrong is judged harshly. One wrong fact is enough to ruin a career.

Theories have an entirely different status. They are free creations of the human mind, intended to describe our understanding of nature. Since our understanding is incomplete, theories are provisional. Theories are tools of understanding; and a tool does not need to be precisely true in order to be useful. Theories are supposed to be more-or-less true, with plenty of room for disagreement. A scientist who invents a theory that turns out to be wrong is judged leniently. Mistakes are tolerated, so long as the culprit is willing to correct them when nature proves them wrong.

—Physicist Freeman Dyson (2014, p. 4)
Institute for Advanced Studies, Princeton

Enhanced Credibility and Increased Legitimacy for Qualitative Methods: Looking Back and Looking Ahead

> **The distinction between the past, present, and future is only a stubbornly persistent illusion.**
>
> —Physicist Albert Einstein
> *Theory of Relativity*

Chapter Summary

This chapter has reviewed ways of enhancing the quality, credibility, and utility of qualitative analysis by dealing with four distinct but related inquiry concerns:

- Rigorous methods for doing fieldwork that yield high-quality data
- Systematic and conscientious analysis with attention to issues of credibility
- The credibility of the researcher, which depends on training, experience, track record, status, and presentation of self
- Philosophical belief in the value of qualitative inquiry—that is, a fundamental appreciation of naturalistic inquiry, qualitative methods, inductive analysis, purposeful sampling, and holistic thinking.

Exhibit 9.15 presented a graphic depicting these four dimensions, with criteria for judging quality in the center.

Conclusion: Beyond the Qualitative/Quantitative Debate

Question: *What's the status of the qualitative/quantitative debate today? From your perspective, what does the future look like for qualitative inquiry?*

The debate between qualitative and quantitative methodologists was often strident historically, but in recent years the debate has mellowed. A consensus has gradually emerged that the important challenge is to appropriately match methods to purposes and inquiry questions, not to universally and unconditionally advocate any single methodological approach for all inquiry situations. Indeed, eminent methodologist Thomas Cook, one of evaluation's luminaries, pronounced in his keynote address to the 1995 International Evaluation Conference in Vancouver that "qualitative researchers have won the qualitative/quantitative debate."

Won in what sense?

Won acceptance.

The validity of experimental methods and quantitative measurement, appropriately used, was never in doubt. Now, qualitative methods have ascended to a level of parallel respectability. I have found increased interest in and acceptance of qualitative methods in particular and multiple methods in general. Especially in evaluation, a consensus has emerged that researchers and evaluators need to know and use a variety of methods in order to be responsive to the nuances of particular empirical questions and the idiosyncrasies of specific stakeholder needs. The debate has shifted from quantitative versus qualitative to strong differences of opinion about how to establish causality (the attribution and so-called gold standard debate discussed in Chapters 3 and 8). While related, that's a narrower issue.

The credibility and respectability of qualitative methods varies across disciplines, university departments, professions, time periods, and countries. In the field I know best, program evaluation, the increased legitimacy of qualitative methods is a function of more examples of useful, high-quality evaluations employing qualitative methods and an increased commitment to providing useful and understandable information based on stakeholders' concerns. Other factors that contribute to increased credibility include more and higher-quality training in qualitative methods and the publication of a substantial qualitative literature.

The history of the paradigms debate parallels the history of evaluation. The earliest evaluations focused largely on quantitative measurement of clear, specific goals and objectives. With the widespread social and educational experimentation of the 1960s and early 1970s, evaluation designs were aimed at comparing the effectiveness of different programs and treatments through rigorous controls and experiments. This was the period when the quantitative/experimental paradigm dominated. By the middle 1970s, the paradigms debate had become a major focus of evaluation discussions and writings. By the late 1970s, the alternative qualitative/naturalistic paradigm had been fully articulated (Guba, 1978; Patton, 1978; Stake, 1975, 1978). During this period, concern about finding ways to increase use became predominant in evaluation, and evaluators began discussing standards. A period of pragmatism and dialogue followed, during which calls for and experiences with multiple methods and a synthesis of paradigms became more common. The advice of Cronbach (1980), in his important book on reform of program evaluation, was widely taken to heart: "The evaluator will be wise not to declare allegiance to either a quantitative–scientific–summative

methodology or a qualitative–naturalistic–descriptive methodology" (p. 7).

Signs of detente and pragmatism now abound. Methodological tolerance, flexibility, eclecticism, and concern for appropriateness rather than orthodoxy now characterize the practice, literature, and discussions of evaluation. Several developments seem to me to explain the withering of the methodological paradigms debate.

1. The articulation of professional standards has emphasized methodological appropriateness rather than paradigm orthodoxy (Joint Committee, 2010; OECD-DAC, 2010). Within the standards as context, the focus on conducting evaluations that are useful, practical, ethical, accurate, and accountable have reduced paradigms polarization.
2. The strengths and weaknesses of both quantitative/experimental methods and qualitative/naturalistic methods are now better understood. In the original debate, quantitative methodologists tended to attack some of the worst examples of qualitative evaluations while the qualitative evaluators tended to hold up for critique the worst examples of quantitative/experimental approaches. With the accumulation of experience and confidence, exemplars of both qualitative and quantitative approaches have emerged with corresponding analyses of the strengths and weaknesses of each. This has permitted more balance and a better understanding of the situations for which various methods are most appropriate as well as grounded experience in how to combine methods.
3. A broader conceptualization of evaluation, and of evaluator training, has directed attention to the relation of methods to other aspects of evaluation, like use, and has therefore reduced the intensity of the methods debate as a topic unto itself.
4. Advances in methodological sophistication and diversity within both paradigms have strengthened diverse applications to evaluation problems. The proliferation of books and journals in evaluation, including but not limited to methods contributions, has converted the field into a rich mosaic that cannot be reduced to quantitative versus qualitative in primary orientation. Moreover, the upshot of all the developmental work in qualitative methods is that, as documented in Chapter 3, today there is as much variation among qualitative researchers as there is between qualitatively and quantitatively oriented scholars.
5. Support for methodological eclecticism from major figures and institutions in evaluation increased methodological tolerance. When eminent measurement and methods scholars like Donald Campbell and Lee J. Cronbach began publicly recognizing the contributions that qualitative methods could make, the acceptability of qualitative/naturalistic approaches was greatly enhanced. Another important endorsement of multiple methods came from the Program Evaluation and Methodology Division of the U.S. General Accounting Office (GAO), which arguably did the most important and influential evaluation work at the national level. Under the leadership of Assistant Comptroller General and former AEA president (1995) Eleanor Chelimsky, GAO published a series of methods manuals, including *Case Study Evaluations* (GAO, 1987), *Prospective Evaluation Methods* (GAO, 1989), and *The Evaluation Synthesis* (GAO, 1992). The GAO manual *Designing Evaluations* put the paradigms debate to rest as it described what constituted a "strong evaluation."

> Strength is not judged by adherence to a particular paradigm. It is determined by use and technical adequacy, whatever the method, within the context of purpose, time, and resources.
>
> Strong evaluations employ methods of analysis that are appropriate to the question; support the answer with evidence; document the assumptions, procedures, and modes of analysis; and rule out competing evidence. Strong studies pose questions clearly, address them appropriately, and draw inferences commensurate with the power of the design and the availability, validity, and reliability of the data. Strength should not be equated with complexity. Nor should strength be equated with the degree of statistical manipulation of data. Neither infatuation with complexity nor statistical incantation makes an evaluation stronger.
>
> The strength of an evaluation is not defined by a particular method. Longitudinal, experimental, quasi-experimental, before-and-after, and case study evaluations can be either strong or weak. . . . That is, the strength of an evaluation has to be judged within the context of the question, the time and cost constraints, the design, the technical adequacy of the data collection and analysis, and the presentation of the findings. A strong study is technically adequate and useful—in short, it is high in quality. (GAO, 1991, pp. 15–16)

6. Evaluation professional societies have supported exchanges of views and high-quality professional practice in an environment of tolerance and eclecticism. The evaluation professional societies

and journals serve a variety of people from different disciplines who operate in different kinds of organizations at different levels, in and out of the public sector, and in and out of universities. This diversity, and opportunities to exchange views and perspectives, has contributed to the emergent pragmatism, eclecticism, and tolerance in the field. A good example was the appearance two decades ago of a volume of New Directions for Program Evaluation on *The Qualitative–Quantitative Debate: New Perspectives* (Reichardt & Rallis, 1994). The tone of the eight distinguished contributions in that volume is captured by phrases such as "peaceful coexistence," "each tradition can learn from the other," "compromise solution," "important shared characteristics," and "a call for a new partnership."

7. There is increased advocacy of and experience in combining qualitative and quantitative approaches. The Reichardt and Rallis (1994) volume just cited also included these themes: "blended approaches," "integrating the qualitative and quantitative," "possibilities for integration," "qualitative plus quantitative," and "working together." Exhibit 9.2 presented 10 developments enhancing mixed-methods triangulation (p. 666).

Matching Claims and Criteria

The withering of the methodological paradigms debate holds out the hope that studies of all kinds can be judged on their merits according to the claims they make and the evidence marshaled in support of those claims. The thing that distinguishes the seven sets of criteria for judging quality introduced in this chapter (Exhibit 9.7) is that they support different kinds of claims. Traditional scientific claims, constructivist claims, artistic claims, participatory inquiry claims, critical change claims, systems claims, and pragmatic claims will tend to emphasize different kinds of conclusions with varying implications. In judging claims and conclusions, the validity of the claims made is only partly related to the methods used in the process.

> Validity is a property of knowledge, not methods. No matter whether knowledge comes from an ethnography or an experiment, we may still ask the same kind of questions about the ways in which that knowledge is valid. To use an overly simplistic example, if someone claims to have nailed together two boards, we do not ask if their hammer is valid, but rather whether the two boards are now nailed together, and whether the claimant was, in fact, responsible for that result. In fact, this particular claim may be valid whether the nail was set in place by a hammer, an airgun, or the butt of a screwdriver. A hammer does not guarantee successful nailing, successful nailing does not require a hammer, and the validity of the claim is in principle separate from which tool was used. The same is true of methods in the social behavioral sciences. (Shadish, 1995a, p. 421)

This brings us back to a pragmatic focus on the utility of findings as a point of entry for determining what's at stake in the claims made in a study and therefore what criteria to use in assessing those claims. As I noted in opening this chapter, judgments about credibility and quality depend on criteria. And though this chapter has been devoted to ways of enhancing quality and credibility, all such efforts ultimately depend on the willingness of the inquirer to weigh the evidence carefully and be open to the possibility that what has been learned most from a particular inquiry is how to do it better next time.

Canadian-born bacteriologist Oswald Avery, discoverer of DNA as the basic genetic material of the cell, worked for years in a small laboratory at the hospital of the Rockefeller Institute in New York City. Many of his initial hypotheses and research conclusions turned out, on further investigation, to be wrong. His colleagues marveled that he never turned argumentative when findings countered his predictions and never became discouraged. He was committed to learning and was often heard telling his students, "Whenever you fall, pick up something."

A final Halcolm story on the nature of journeys ends this chapter—and this book.

EXHIBIT 9.16 A Documenter's Perspective

by Beth Alberty

Introduction

This exhibit provides a reflective case study of the struggle experienced by one internal, formative program evaluator of an innovative school art program as she tried to figure out how to provide useful information to program staff from the voluminous qualitative data she collected. Beth begins by describing what she means by "documentation" and then shares her experiences as a novice in analyzing the data, a process of moving from a mass of documentary material to a unified, holistic document.

Documentation

Documentation, as the word is commonly used, may refer to "slice of life" recordings in various media or to the marshalling of evidence in support of a position or point of view. We are familiar with "documentary" films; we require lawyers or journalists to "document" their cases. Both meanings contribute to my view of what documentation is, but they are far from describing it fully. Documentation, to my mind, is the interpretive reconstitution of a focal event, setting, project, or other phenomenon, based on observation and on descriptive records set in the context of guiding purposes and commitments.

I have always been a staff member of the situations I have documented, rather than a consultant or an employee of an evaluation organization. At first this was by accident, but now it is by conviction: My experience urges that the most meaningful evaluation of a program's goals and commitments is one that is planned and carried out by the staff and that such an evaluation contributes to the program as well as to external needs for information. As a staff member, I participate in staff meetings and contribute to decisions. My relationships with other staff members are close and reciprocal. Sometimes I provide services or perform functions that directly fulfill the purposes of the program—for example, working with children or adults, answering visitor's questions, and writing proposals and reports. Most of my time, however, is spent planning, collecting, reporting, and analyzing documentation.

First Perceptions

With this context in mind, let me turn to the beginning plunge. Observing is the heart of documenting, and it was into observing that I plunged, coming up delighted at the apparent ease and swiftness with which I could fish insight and ideas from the ceaseless ocean of activity around me. Indeed, the fact that observing (and record keeping) does generate questions, insight, and matters for discussion is one of many reasons why records for any documentation should be gathered by those who actually work in the setting.

My observing took many forms, each offering a different way of releasing questions and ideas—interactive and noninteractive observations were transcribed or discussed with other staff members and thereby rethought; children's writing was typed out, the attention to every detail involving me in what the child was saying; notes of meetings and other events were rewritten for the record; and so on. Handling such detail with attention, I found, enabled me to see into the incident or piece of work in a way I hadn't on first look. Connections with other things I knew, with other observations I made, or questions I was puzzling over seemed to proliferate during these processes; new perceptions and new questions began to form.

I have heard others describe similarly their delighted discovery of the provocativeness of record-keeping processes. The teacher who begins to collect children's art, without perhaps even having a particular reason for the collecting, will, just by gathering the work together, begin to notice things about them that he or she had not seen before—how one child's work influences another's, how really different (or similar) are the trees they make, and so on. The in-school advisor or resource teacher who reviews all his or her contacts with teachers—as they are recorded or in a special meeting with his or her colleagues—may begin, for example, to see patterns of similar interest in the requests he or she is getting and thus become aware of new possibilities for relationships within the school.

My own delight in this apparently easy access to a first level of insight made me eager to collect more and more, and I also found the sheer bulk of what I could collect satisfying. As I collected more records, however, my enthusiasm gradually changed to alarm and frustration. There were so many things that could be observed and recorded, so many perspectives, such a complicated history! My feelings of wanting more changed to a feeling of needing to get everything. It wasn't enough for me to know how the program worked now—I felt I needed to know how it got started and how the present workings had evolved. It wasn't enough to know how the central part of the program worked—I felt I had to know about all its spinoff activities and from all points of view. I was quickly drawn into a fear of losing something significant,

something I might need later on. Likewise, in my early observations of class sessions, I sought to write down everything I saw. I have had this experience of wanting to get everything in every setting in which I have documented, and I think it is not unique.

I was fortunate enough to be able to indulge these feelings and to learn from where they led me. It did become clear to me after a while that my early ambitions for documenting everything far exceeded my time and, indeed, the needs of the program. Nevertheless, there was a sense to them. Collecting so much was a way of getting to know a new setting, of orienting myself. And, not knowing the setting, I couldn't know what would turn out to be important in "reconstituting" it; also, the purpose of "reconstituting" it was sufficiently broad to include any number of possibilities from which I had not yet selected. In fact, I found that the first insights, the first connections that came from gathering the records were a significant part of the process of determining what would be important and what were the possibilities most suited to the purposes of the documentation. The process of gathering everything at first turned out to be important and, I think, needs to be allowed for at the beginning of any documenting effort. Even though much of the material so gathered may remain apparently unused, as it was in my documenting, in fact it has served its purpose just in being collected. A similar process may be required even when the documenter is already familiar with the setting, since the new role entails a new perspective.

The first connections, the first patterns emerging from the accumulating records were thus a valuable aspect of the documenting process. There came a moment, however, when the data I had collected seemed more massive than was justified by any thought I'd had as a result of the collecting. I was ill at ease because the first patterns were still fairly unformed and were not automatically turning into a documentation in the full sense I gave earlier, even though I recognized them as part of the documentary data. Particularly, they did not function as "evaluation." Some further development was needed, but what? "What do I do with them now?" is a cry I have heard regularly since then from teachers and others who have been collecting records for a while.

I began with the relatively simple procedure of rereading everything I had gathered. Then, I returned to rethink what my purposes were and sought out my original resources on documentation. Rereading qualitative references, talking with the staff of the school and with my staff colleagues, I began to imagine a shape I could give to my records that would make a coherent representation of the program to an outside audience.

At the same time, I began to rethink how I could make what I had collected more useful to the staff. Conceiving an audience was very important at this stage. I will be returning to this moment of transition from initial collecting to rethinking later, to analyze the entry into interpretation that it entails. Descriptively, however, what occurred was that I began to see my observations and records as a body with its own configurations, interrelationships, and possibilities, rather than simply as excerpts of the larger program that related only to the program. Obviously, the observations and records continued to have meaning through their primary relationship to the setting in which they were made; but they also began to have meaning through their secondary relationships to each other.

These secondary relationships also emerge from observation as a process of reflecting. Here, however, the focus of observation is the setting as it appears in and through the observations and records that have accumulated, with all their representation of multiple perspectives and longitudinal dimensions. These observations in and through records —"thickened observations"—are of course confirmed and added to by continuing direct observation of the setting.

Beginning to see the records as a body and the setting through thickened observation is a process of integrating data. The process occurs gradually and requires a broad base of observation about many aspects of the program over some period of time. It then requires concentrated and systematic efforts to find connections within the data and weave them into patterns, to notice changes in what is reported, and find the relationship of changes to what remains constant. This process is supported by juxtaposing the observations and records in various ways as well as by continual return to reobserve the original phenomenon. There is, in my opinion, no way to speed up the process of documenting. Reflectiveness takes time.

In retrospect, I can identify my own approach to an integration of the data as the time when I began to give my opinions on long-range decisions and interpretations of daily events with the ease of any other staff member. Up to the moment of transition, I shared specific observations from the records and talked them over as a way of gathering yet more perspectives on what was happening. I was aware, however, that my opinions or interpretations were still personal. They did not yet represent the material I was collecting.

(Continued)

(Continued)

Thus, it may be that integration of the documentary material becomes apparent when the documenter begins to evince a broad perspective about what is being documented, a perspective that makes what has been gathered available to others without precluding their own perceptions. This perspective is not a fixed-point view of a finished picture, both the view and the picture constructed somehow by the documenter in private and then unveiled with a flourish. It is also not a personal opinion; nor does it arise from placing a predetermined interpretive structure or standard on the observations. The perspective results from the documenter's own current best integration of the many aspects of the phenomenon, of the teachers' or staff's aims, ideas, and current struggles, and of their historical development as these have been conveyed in the actions that have been observed and the records that have been collected.

As documenter, my perspective of a program or a classroom is like my perspective of a landscape. The longer I am in it, the sharper defined become its features, its hills and valleys, forests and fields, and the folds of distance; the more colorful and yet deeply shaded and nuanced in tone it appears; the more my memory of how it looks in other weather, under other skies, and in other seasons, and my knowledge of its living parts, its minute detail, and its history deepen my viewing and valuing of it at any moment. This landscape has constancy in its basic configurations, but is also always changing as circumstances move it and as my perceptions gather. The perspective the documenter offers to others must evoke the constancy, coherence, and integrity of the landscape, and its possibilities for changing its appearance. Without such a perspective, an organization or integration that is both personal and informed by all that has been gathered by myself and by others in the setting—others could not share what I have seen—could not locate familiar landmarks and reflect on them as they exhibit new relationships to one another and to less familiar aspects. All that material, all those observations and records, would be a lifeless and undoubtedly dusty pile.

The process of forming a perspective in which the data gathered are integrated into an organic configuration is obviously a process of interpretation. I had begun documenting, however, without an articulated framework for interpretation or a format for representation of the body of records, like the theoretical framework researchers bring to their data. Of course, there was a framework. Conceptions of artistic process, of learning and development, were inherent in the program; but these were not explicit in its goals as a program to provide certain kinds of service. The plan of the documentation had called for certain results, but there was no specified format for presentation of results. Therefore, my entry into interpretation became a struggle with myself over what I was supposed to be doing. It was a long internal debate about my responsibilities and commitments.

When I began documenting this particular school's art program, for example, I had priorities based on my experience and personal commitments. It seemed to me self-evidently important to provide art activities for children and to try and connect these to other areas of their learning. I knew that art was not something that could be "learned" or even experienced on a once-a-week basis, so I thought it was important to help teachers find various ways of integrating art and other activities into their classrooms. I had already made a personal estimate that what I was documenting was worthwhile and honest. I had found points of congruence between my priorities and the program. I could see how the various structures of the program specified ways of approaching the goals that seemed possible and that also enabled the elaboration of the goals.

This initial commitment was diffuse; I felt a kind of general enthusiasm and interest for the efforts I observed and a desire to explore and be helpful to the teachers. In retrospect, however, the commitment was sufficiently energizing to sustain me through the early phases of collecting observations and records, when I was not sure what these would lead to. Rather than restricting me, the commitment freed me to look openly at everything (as reflected in the early enthusiasm for collecting everything). Obviously, it is possible to begin documenting from many other positions of relative interest and investment, but I suspect that even if there is no particular involvement in program content on the part of the documenter, there must be at least some idea of being helpful to its staff. (Remember, this was a formative evaluation.) Otherwise, for example, the process of gathering data may be circumscribed.

At the point of beginning to "do something" with the observations and records, I was forced to specify the original commitment, to rethink my purposes and goals. Rereading the observations and records as a preliminary step in reworking to address different audiences, I found myself at first reading with an idea of "balancing" success and failure, an idea that constricted and trivialized the work I had observed and recorded. Thankfully, it was immediately evident from the data itself that such balance was not possible. If, during 10 days of observation, a child's experience was intense 1 day and characterized by rowdy socializing the other 9, a simple weigh-off would

not establish the success or failure of the child's experience. The idea was ludicrous. Similarly, the staff might be thorough in its planning and follow-through on one day and disorganized on another day, but organization and planning were clearly not the totality of the experience for children.

Such trade-offs implied an external, stereotyped audience awaiting some kind of quantitative proof, which I was supposed to provide in a disinterested way, like an external, summative evaluator. The "balanced view" phase was also like my early record gathering of everything. What I was documenting was still in fragments for me, and my approach was to the particulars, to every detail.

A second approach to interpreting, also brief, took a slightly broader view of the data, a view that acknowledged my original estimate of program value and attempted to specify it. Perceiving through the data the landscape-like configurations of program strengths, I made assessments that included statements of past mistakes or inadequacies like minor "flaws" in the landscape (e.g., a few odd billboards and a garbage dump in one of Poussin's dreams of classical Italy) rather than debits on a balance sheet. Here again, the implication was of an external audience, expecting some absolute of accomplishment. The "flaws" could be "minor" only by reference to an implied major flaw—that of failing to carry out the program goals altogether.

The formulation of strength subsuming weakness could not withstand the vitality of the records I was reading. The reality the data portrayed became clearer as the inadequacy of my first formulations of how to interpret the documentary material was revealed. Similarly, the implications of external audience expectations were not justified by the actuality of my relationship to the program and staff. My stated goal as documenter had been originally to set up record-keeping procedures that would preserve and make available to staff and to other interested persons aspects of the beginnings and workings of the program, and to collect and analyze some of the material as an assessment of what further possibilities for development actually existed. My goals had not been to evaluate in the sense of an external judgment of success or failure.

Thinking over what other approaches to interpretation were possible, I recalled that I had gathered documentary materials quite straightforwardly as a participant, whose engagement was initially through recognition of shared convictions and points of congruence with the program. Perhaps, I decided, I could share my viewpoint of the observations just as straightforwardly, as a participant with a particular point of view. In examining this possibility, I came to a view of interpreting observational data as a process of "rendering," much as a performer renders a piece of classical music. The interpretation follows a text closely—as a scientist might say, it sticks closely to the facts. But it also reflects the performer, specifically the performer's particular manner of engagement in the enterprise shared by text and performer, the enterprise of music. The same relationship could exist, it seemed to me, between a body of observations and records gathered participatively and as documenter. The relationship would allow my personal experience and viewpoint to enhance rather than distort the data. Indeed, I would become their voice.

Through this relationship I could make the observations available to staff and to other audiences in a way that was flexible and responsive to *their* needs, purposes, and standards. In so doing, of course, the framework of inherent conceptions underlying the work of the program would be incorporated. Thus, to interpret the observational data I had gathered, I had to reaffirm and clarify my relationship, my attachment to and participation in the program.

My initial engagement, with its strong coloring of prior interests and ideas, had never meant that I understood or was sympathetic with every goal or practice of every participant of the program all the time. In any joint enterprise, such as a school or program, there are diverse and multiple goals and practices. Part of the task of documenting is to describe and make these various understandings, points of view, and practices visible so that participants can reflectively consider them as the basis for planning. No participant agrees on all issues and points of practice. Part of being a participant is exploring differences and how these illuminate issues or contribute to practice. My participation allowed me to examine and extend the interests and ideas I came with as well as observing and recording those other people brought. In this process, my engagement was deepened, enabling me to make assessments closer to the data than my first readings brought. These assessments are evaluation in its original sense of "drawing-value-from," an interactive process of valuing, of giving weight and meaning.

In the context of renewed engagement and deepened participation, assessments of mistakes or inadequacies are construed as discrepancies between a particular practice and the intent behind it, between immediate and long-range purposes. The discrepancy is not a flaw in an otherwise perfect surface, but—like the discrepancy in a child's understanding that stimulates new learning—is the

(Continued)

(Continued)

occasion for growth. It is a sign of life and possibility. The burden of the discrepancy can lie either with the practice or with the intent, and that is the point for further examination. Assessment can also occur through the observation of and search for underlying themes of continuity between present and past intent and practice, and the point of change or transformation in continuity. Whereas discrepancy will usually be a more immediate trigger to evaluation, occasions for the consideration of continuity may tend to be longer-range planning for the coming year, contemplating changes in staff and function, or commemorating an anniversary.

I have located the documenter as participant, internal to the program or setting, gathering and shaping data in ways that make them available to participants and potentially to an external audience. Returning to the image of a landscape, let me comment on the different forms availability assumes for these different audiences.

Participant access to the landscape through the documenter's perspective cannot be achieved through ponderous written descriptions and reports on what has been observed but must be concentrated in interaction. Sometimes this may require the development of special or regular structures—a series of short-term meetings on a particular issue or problem; an occasional event that sums up and looks ahead; a regular meeting for another kind of planning. But many times the need is addressed in very slight forms, such as a comment in passing about something a child or adult user is doing, about the appearance of a display, or the recounting of another staff member's observation. I do not mean that injecting documentation into the self-assessment process is a juggling act or some feat of manipulation; merely that the documenter must be aware that his or her role is to keep things open and that, while the observations and records are a resource for doing this, a sense of the whole they create is also essential. The landscape is, of course, changed by the new observations offered by fellow viewers.

The external audience places different requirements on the documenter who seeks to represent to it the documentary perspective. By external audience I refer to funding agencies, supervisors, school boards, institutional hierarchies, and researchers. Proposals, accounts, and reports to these audiences are generally required. They can be burdensome because they may not be organically related to the process of internal self-reflection and because the external audience has its own standards, purposes, and questions; it is unfamiliar with the setting and with the documenter, and it needs the time offered by written accounts to return and review the material. The external audience will need more history and formal description of the broad aspects than the internal audience, with commentary that indicates the significance of recent developments. This need can be met in the overall organization, arrangement, and introduction of documents, which also convey the detail and vividness of daily activity.

To limit the report to conventional format and expectations would probably misrepresent the quality of thought, of relating, of self-assessment that goes into developing the work. If there is intent to use the occasion of a report for reflection—for example, by including staff in the development of the report—the reporting process can become meaningful internally while fulfilling the legitimate external demands for accounting. Naturally, such a comment engages the external audience in its own evaluative reflections by evoking the phenomenon rather than reducing it.

In closing, I return to what I see as the necessary engaged participation of the documenter in the setting being documented, not only for data gathering but for interpretation. Whatever authenticity and power my perspective as documenter has had has come, I believe, from my commitment to the development of the setting I was documenting and from the opportunities in it for me to pursue my own understanding, to assess and reassess my role, and to come to terms with issues as they arose.

We come to new settings with prior knowledge, experience, and ways of understanding, and our new perceptions and understandings build on these. We do not simply look at things as if we had never seen anything like them before. When we look at a cluster of light and dark greens with interstices of blue and some of deeper browns and purples, what we identify is a tree against the sky. Similarly, in a classroom we do not think twice when we see, for example, a child scratching his head, yet the same phenomenon might be more strictly described as a particular combination of forms and movements. Our daily functioning depends on this kind of apparently obvious and mundane interpretation of the world. These interpretations are not simply personal opinions—though they certainly may be unique—nor are they made up. They are instead organizations of our perceptions as "tree" or "child scratching" and they correspond at many points with the phenomena so described.

It is these organizations of perception that convey to someone else what we have seen and that make objects available for discussion and reflection. Such organizations need not exclude our awareness that the tree is also a cluster of colors or that the child scratching his head is

also a small human form raising its hand in a particular way. Indeed, we know that there could be many other ways to describe the same phenomena, including some that would be completely numerical—but not necessarily more accurate, more truthful, or more useful! After all, we organize our perceptions in the context of immediate purposes and relationships. The organizations must correspond to the context as well as to the phenomenon.

Facts do not organize themselves into concepts and theories just by being looked at; indeed, except within the framework of concepts and theories, there are no scientific facts but only chaos. There is an inescapable a priori element in all scientific work. Questions must be asked before answers can be given. The questions are all expressions of our interest in the world; they are at bottom valuations. Valuations are thus necesssarily involved already at the stage when we observe facts and carry on theoretical analysis and not only at the stage when we draw political inferences from facts and valuations (Myrdal, 1969, p. 9).

My experience suggests that the situation in documenting is essentially the same as what I have been describing with the tree and the child scratching and what Myrdal describes as the process of scientific research. Documentation is based on observation, which is always an individual response both to the phenomena observed and to the broad purposes of observation. In documentation, observation occurs both at the primary level of seeing and recording phenomena and at secondary levels of re-observing the phenomena through a volume of records and directly, at later moments. Since documentation has as its purpose to offer these observations for reflections and evaluation in such a way as to keep alive and open the potential of the setting, it is essential that observations at both primary and secondary levels be interpreted by those who have made them. The usefulness of the observations to others depends on the documenter's rendering them as finely as he or she is able, with as many points of correspondence to both the phenomena and the context of interpretation as possible. Such a rendering will be an interpretation that preserves the phenomena and so does not exclude but rather invites other perspective.

Of course, there is a role for the experienced observer from outside who can see phenomenon freshly; who can suggest ways of obtaining new kinds of information about it, or, perhaps more important, point to the significance of already existing procedures or data; who can advise on technical problems that have arisen within a documentation; and who can even guide efforts to interpret and integrate documentary information. I am stressing, however, that the outside observer in these instances provides support, not judgment or the criteria for judgment.

The documenter's obligation to interpret his or her observations and those reflected in the records being collected becomes increasingly urgent, and the interpretations become increasingly significant, as all the observers in the setting become more knowledgeable about it and thus more capable of bringing range and depth to the interpretation. Speaking of the weight of her observations of the Manus over a period of some 40 years to great change, Margaret Mead clarifies the responsibility of the participant–observer to contribute to both people studied and to a wider audience the rich individual interpretation of his or her own observations:

> Uniqueness, now, in a study like this (of people who have come under the continuing influence of contemporary world culture), lies in the relationships between the fieldworker and the material. I still have the responsibility and incentives that come from the fact that because of my long acquaintance with this village I can perceive and record aspects of this people's life that no one else can. But even so, this knowledge has a new edge. This material will be valuable only if I myself can organize it. In traditional fieldwork, another anthropologist familiar with the area can take over one's notes and make them meaningful. But here it is my individual consciousness that provides the ground on which the lives of these people are figures. (Mead, 1977, pp. 282–283)

In documenting, it seems to me the contribution is all the greater, and all the more demanded, because what is studied is one's own setting and commitment.

APPLICATION EXERCISES

1. Locate a published qualitative study on a subject of interest to you. How does the study address and establish the credibility of qualitative inquiry? Use Exhibits 9.4 and 9.15 to review and critique how credibility is addressed in the study you've chosen. What questions are left unanswered in the study you're reviewing that, from your perspective, if answered, would enhance credibility?

2. Locate a study that highlights use of mixed methods. What was the nature of the mix? What rationale was used for mixing methods? To what extent was *triangulation* an explicit justification for mixing methods? How integrated was the analysis of qualitative and quantitative data? Based on your review, what are the strengths and weaknesses of the mixed-methods design and analysis you reviewed.

3. Exhibit 9.11 (pp. 707–709) provides the framework for establishing the credibility of a qualitative inquirer. If you have conducted a qualitative study, or been part of one, complete that table using your own experience (fill in the column for yourself that reports Nora Murphy's experiences, perspectives, reactions, and competence in Exhibit 9.11). If you haven't done a qualitative study, imagine one and complete the table for a qualitative scenario that you construct. The purpose is to practice being reflexive and addressing inquirer credibility.

4. The discussion on objectivity considers a number of alternative ways of describing an inquirer's stance and philosophy (pp. 723–728). What is your preferred terminology? Write a statement describing your paradigm stance, a statement that you could give someone who was considering funding you to do a qualitative study. Describe a scenario or situation where you would need to explain your stance—and then do so. (You don't have to be limited to the language options discussed here.)

5. a. As an exercise in distinguishing quality criteria frameworks, try matching the three umpires' perspectives (p. 683) to the frameworks in Exhibit 9.7 (pp. 680–881). Explain your choices.

   b. What would a systems-oriented umpire say about umpiring? (Explain).

   c. What would an artistic-evocative-oriented umpire say about umpiring (Explain).

   d. What would a critical change umpire say?

6. (Advanced application) On the next page is a description of an edited volume of qualitative inquiries into the nature of family. Use the criteria for autoethnography in Chapter 3 (pp. 102–104) and the sets of criteria in Exhibit 9.7. Create your own set of 10 criteria for judging the methodological quality of this book by selecting criteria that seem especially relevant given the description of the book's approach. Use this example to discuss the nature of quality criteria in judging the quality of qualitative inquiries.

   From *On (Writing) Families: Autoethnographies of Presence and Absence, Love and Loss* (Wyatt & Adams, 2014):

> Who are we with—and without—families? How do we relate as children to our parents, as parents to our children? How are parent–child relationships—and familial relationships in general—made and (not) maintained?
>
> Informed by narrative, performance studies, poststructuralism, critical theory, and queer theory, contributors to this collection use autoethnography—a method that uses the personal to examine the cultural—to interrogate these questions. The essays write about/around issues of interpersonal distance and closeness, gratitude and disdain, courage and fear, doubt and certainty, openness and secrecy, remembering and forgetting, accountability and forgiveness, life and death.
>
> Throughout, family relationships are framed as relationships that inspire and inform, bind and scar—relationships replete with presence and absence, love and loss (p. 1).

7. (Advanced application) Martin Rees, astronomer, former Master of Trinity College, and ex-president of the Royal Society of Astronomy said, "Ultimately, I don't think there's anything special in the scientific method that goes beyond what a detective does" (quoted by Morris, 2014). Imagine that you are using this quotation to support the credibility of qualitative inquiry. Discuss how this quotation applies to each of the four dimensions of credibility discussed in this chapter (see Exhibit 9.15, p. 722)