QUALITATIVE RESEARCH METHODS IN PROGRAM EVALUATION: CONSIDERATIONS FOR FEDERAL STAFF

Office of Data, Analysis, Research & Evaluation
Administration on Children, Youth & Families

May 2016

Table of Contents

EXECUTIVE SUMMARY

Qualitative research methods can play a powerful role in program evaluation, but they frequently are misunderstood and poorly implemented, giving rise to the idea that they are just not as rigorous and credible as quantitative methods. That is not true, but qualitative methods are less standardized than are quantitative methods, and that makes determining their appropriate use and assessing their quality more difficult for federal staff overseeing program evaluation projects that include a qualitative component.

This document was written to support federal program officers, particularly those who have not had formal research training, as they develop and oversee projects that include program evaluations that use qualitative methods. Divided into four parts, this paper begins with the decision to include qualitative methods (or not) in a contemplated project. This first section defines qualitative methods, distinguishes them from quantitative research methods, and outlines the roles qualitative approaches can play in program evaluation. This portion of the document will be particularly useful as a project or a funding announcement is being developed.

The second section of the paper takes up a couple of early considerations, once the decision to incorporate qualitative methods has been made: How to identify an evaluator with the requisite experience in these methods and addressing some time and cost considerations that are particular to qualitative research. These points will be helpful both in drafting the funding announcement and in reviewing applications. The third section of the paper provides a brief overview of qualitative research methods addressing research design, sampling, data collection, and data analysis. The idea is to provide some familiarity with concepts and terminology that will help federal staff communicate more effectively with evaluators as the project progresses; this section also may be of use in evaluating proposals that include program evaluation.

Finally, the document takes up how to assess the credibility of qualitative findings and conclusions, addressing both internal and external validity. Although the goal of obtaining credible findings drives any research project from its inception, it may be easier to follow the discussion of credibility if the reader has received some grounding in qualitative research methods. Throughout this document are questions federal staff can ask themselves or their evaluators to aid decision-making as well as to press evaluators to be explicit about their methodological choices, time-planning, and budgeting.

INTRODUCTION

This document has been prepared for staff charged with overseeing federally-funded program evaluations that include the collection and analysis of qualitative data. The information is organized and presented in four parts to allow you to get quickly to the information you need, depending on your level of expertise and/or the project's stage of work — developing the funding announcement, reviewing proposals or evaluation plans, overseeing the work as it progresses, and approving final reports and other products.

Not all federal staff have formal research training; the intent of this document is to equip you with a basic knowledge of qualitative methods and when they may be most useful, how to determine the quality of the findings, and to help you consider the trade-offs involved in deciding when to use these methods. It also should help dispel the oft-expressed idea that qualitative methods are not as rigorous as quantitative research. That simply is not true, although it has been the case the qualitative research has not always been well-executed, and poor execution certainly compromises the utility and credibility of the findings. As noted in ACF's Evaluation Policy, "Rigor is not restricted to impact evaluations, but is also necessary in implementation or process evaluations, descriptive studies, outcome evaluations, and formative evaluations; and in both qualitative and quantitative approaches."[1] This document will give you some grounding in what good qualitative research looks like and how much careful, rigorous work goes into a well-executed project. This document also suggests questions you can ask of evaluators to help you in your decision-making, as well as to press them to be explicit about their methodological choices and to deliver a usable, credible product.

The document is organized as follows: Part One lays the foundation by defining exactly what qualitative methods are, the kinds of questions that can best be addressed by this approach, and the role qualitative methods can play in program evaluation. Just as quantitative methods are well suited for some types of questions, qualitative methods are particularly well suited to other, specific types of questions. Therefore, the first step is to consider what questions the study needs to answer and whether qualitative methods are appropriate for answering them. The information in this section should be helpful for deciding whether or not to include a request for qualitative research in a funding announcement, as well as for wording the announcement and reviewing proposals.

Part Two takes up a couple of early-stage steps, once the decision to incorporate qualitative methods has been made: the importance of engaging an evaluator with specific experience in using qualitative methods and a consideration of the time and budget implications of the decision to use these approaches. This information should be of use for the development of the funding announcement and in reviewing applications.

Part Three provides a high-level overview of qualitative research methods, including research design, sampling, data collection, and data analysis. It also covers methodological considerations attendant upon research fieldwork: researcher bias and data collection by program staff. This section will familiarize you with terminology and concepts, as well as provide a sense of what, exactly, researchers do when they collect and analyze qualitative data. The goal is to equip you to communicate effectively with evaluators and project directors. This

[1] Administration for Children and Families, Evaluation Policy. 2012. http://www.acf.hhs.gov/programs/opre/resource/acf-evaluation-policy

section should be useful at the proposal review stage, when approving evaluation plans, and in monitoring work as it progresses.

Part Four dives a bit deeper into method — particularly analysis and interpretation — to discuss what determines the validity and credibility of qualitative findings and conclusions. The goal of generating valid, credible findings should guide the project at all stages of the work, so the presence of this section at the end of the document in no way suggests that it applies only at the conclusion of a project. Not at all. From the selection of an experienced investigator, though the implementation of an appropriate research design and sampling strategy, and the careful analysis of the data, credibility and validity should be the project's guide-stars. That said, this final section will also be useful during the later stages of the project for thinking about the credibility of the findings and how the conclusions can be extended to other settings. However, it is too late in the game to find out, only as the final report is written, that the study's findings cannot be relied upon. Therefore, looking this section over, even as a project gets started, will help you to start thinking about how to ensure a quality product.

PART ONE: DECIDING WHEN TO USE QUALITATIVE METHODS

References to program evaluation methods frequently include the phrase "qualitative and quantitative methods," as if the mention of one method demands the inclusion of the other. Although methodological diversity in evaluation is widely accepted, and even recommended by some observers, it remains necessary to consider carefully the goals of any given program evaluation and to select the approach most suitable for answering the questions at hand, rather than reflexively calling for both (Patton, 2002; Schorr & Farrow, 2011). Depending on the research questions to be answered, the best approach may be quantitative, qualitative, or some combination of the two.

What follows is intended to help you decide what benefits, if any, qualitative data can provide to a given project. Even though qualitative data often are collected under less structured research designs and on a comparatively small sample of subjects, the enormous amount of data generated — and the time and expertise needed to collect, organize, and analyze those data — means that qualitative studies are at least as expensive, and can be more costly, than quantitative studies (Morse, 2003).[2] Therefore, it is essential to be clear about when qualitative techniques are called for and to be prepared to fund the project adequately to ensure credible, high quality data.

The following discussion begins by defining exactly what qualitative methods are and how they differ from quantitative research. Next, the particular strengths of qualitative methods within program evaluation are discussed.

What are Qualitative Methods?

Let us start by clarifying exactly what qualitative methods are — and what they are not. The broad term "methods" is used to apply to the collection, analysis, interpretation, and presentation of research data. This brief will address methods used with qualitative data as these differ from those used for quantitative data. Typically gathered in the field, that is, the setting being studied, qualitative data used for program evaluation are obtained from three sources (Patton, 2002):

- **In-depth interviews that use open-ended questions:** "Interviews" include both one-on-one interviews and focus groups. "Open-ended" questions are those that allow the informants to express responses using their own words. These questions may be embedded in interviews that are structured, unstructured, or semi-structured; the open-endedness is what makes the interview qualitative in nature.[3]

[2] The next section of this document touches briefly on time and budget matters related to qualitative methods.

[3] It is important not to conflate question type with interview structure: Structured, semi-structured, and unstructured interviews, which are defined in the data collection section, are characterized by how rigidly the interviewer has to adhere to a pre-defined interview protocol. These terms do not apply to the type of questions (i.e., open-ended or forced choice) included in those interviews. That said, qualitative interviews, structured or otherwise, may include a limited number of forced-choice questions.

- **Direct observation** yields detailed descriptions of the activities, actions, and behaviors of individuals; interpersonal interactions; settings; and organizational processes and procedures.
- **Document analysis** may include the full range of organizational, programmatic, or clinical records, including public reports, memoranda, policy documents, correspondence, and the like.

Quantitative methods may also use some of these data collection approaches; the difference between quantitative and qualitative is in how the data are captured and expressed. In quantitative research, data are expressed numerically. In contrast, qualitative data most often are in the form of words — interview or focus group transcripts, observational field notes, or excerpts from documents (Miles & Huberman, 1994; National Research Council and Institute of Medicine, 2002). Analysis of such data consists of extracting themes, patterns, categories, and case examples (Patton, 2002; Hood, 2006). The purpose of qualitative analysis is to understand how people involved with the program being studied understand, think about, make sense of, and manage situations in their lives and environment and/or to describe the social or environmental contexts within which a program is implemented.

Despite the differences between qualitative and quantitative methods, the line between the two is less distinct than it may seem. For example, although the presentation of qualitative findings relies primarily on words and focuses on patterns and themes, quantitative concepts also may be expressed. The presentation of qualitative findings, in addition to in-text descriptions, may include counts of how many respondents expressed a particular theme, sometimes presented in a table or matrix, and the report text often will include words like "often" or "rarely," which express quantitative concepts (Secrest & Sidani, 1995). However, quantitative terms and concepts serve mainly to organize and summarize qualitative findings. The focus on detailed description expressed in words and analyzed for meaning is what characterizes qualitative research methods.

The Role of Qualitative Methods in Program Evaluation

In contrast to quantitative methods, which ask variations of "how much/many" questions, qualitative methods focus more on "how" and "why" types of questions (James Bell Associates, 2009). Qualitative inquiry places a priority on people's lived experience and the meanings they ascribe to their experiences (Miles & Huberman, 1994). Data often are collected in the settings under study, and they aim for rich description of complex ideas or processes, albeit typically across a limited number of individuals or settings. This approach stands in contrast to quantitative methods, which explore variables that can be captured or represented in numerical form, often across large samples and/or multiple points in time.

Although qualitative methods may be used in both formative and summative evaluations, as a practical matter, they tend to be more heavily relied upon in formative evaluations. Summative evaluations — that is, those that are aimed at determining the effectiveness of the program — often use qualitative methods mainly to add context and detail, while quantitative data play the major role in measuring outcomes (Patton, 2002).

The choice of method — qualitative or quantitative — should always be driven by the research questions to be answered, and evaluations may choose to use both. Given their

particular strengths, qualitative methods can play a number of roles in program evaluation. You may see them proposed for the following research tasks:

- Theory of Change/Logic Model Development: Long regarded as useful in general research for exploring new areas of inquiry and generating hypotheses (Miles & Huberman, 1994), qualitative methods translate well to the analogous stages of program evaluation. Because they can collect detailed information to describe and analyze how programs operate, these methods are useful for formulating or modifying a program's theory of change and the development of a logic model (National Research Council and Institute of Medicine, 2002; Patton, 2002). Logic models are graphic representations of a program's inputs, activities, and short-, medium-, and long-range outcomes (Clark & Anderson, 2004; Office of Planning, Research and Evaluation, 2010). Theories of change, which explain how and why a program exerts its effect on the target population, may be generated through qualitative methods and later can be tested either qualitatively or quantitatively in a comprehensive program of evaluation (Clark & Anderson, 2004; Creswell et al., 2011; Framework Workgroup, 2014; Patton, 2002; Shadish, 1995b).

- When Established Measures Are Inappropriate or Do Not Exist: Because qualitative data collection generally does not rely on standardized measurement instruments, using these approaches in the initial stages of a project, or for circumstances or populations for which no established measures exist, is appropriate (Creswell et al., 2011; Miles and Huberman 1994). For populations who are uncomfortable with quantitative measures, or for whom storytelling and narrative are more familiar methods of conveying information, such as tribal communities, qualitative techniques may be particularly valuable (Tribal Evaluation Workgroup, 2013).

- Studying Program Implementation: The types of questions that implementation science identifies as critical to implementation fidelity — How well defined are practices and programs? What is the level of buy-in or readiness for the program among community members and other stakeholders? How well developed are the program provider's staff hiring, training, coaching, and evaluation practices? What efforts were made to bring about the organizational change necessary to implement the program? — call for detailed, descriptive information (Fixsen et al., 2005). Therefore, qualitative methods are particularly useful for these types of projects, although quantitative methods may also be appropriate. Furthermore, having thorough documentation of how a program was implemented, of the processes believed to bring about client change, and the impact on outcomes of contextual factors, can be invaluable for interpreting outcomes measures or for adapting the program for use in other contexts or with other populations (Fixsen et al., 2005; Framework Workgroup, 2014; National Research Council and Institute of Medicine, 2002; Patton, 2002; Rist, 2000; Schorr and Farrow, 2011).

- Opening the "Black Box" of Program Effects: Measuring differences in outcomes among experimental and control groups — or between treatment and comparison groups in a quasi-experimental design — tells you if the program had an effect, but does not usually tell you why. Qualitative methods are particularly well-suited to explaining why a program had the effect that it did — or failed to have such an effect — by "getting at the

stories" behind the quantitative measures (National Research Council and Institute of Medicine, 2002; Patton, 2002). Adding qualitative approaches to an evaluation study design can shed light on these questions by documenting the experiences of clients and staff, examining contextual changes that might affect outcomes, exploring what the outcomes mean to program participants, or uncovering unintended programmatic side effects (National Research Council and Institute of Medicine, 2002; James Bell Associates, 2009; Miles & Huberman, 1994; Patton, 2002; Puddy & Wilkins, 2011; Rist, 2000). As noted above, documenting the process of implementation also is useful for understanding program effects (Fixsen et al., 2005).

- Making Research Reports More Accessible: When a program evaluation report is meant to be read by a wide variety of audiences — some of whom will not have formal research training — it makes good sense to have at least some of the findings presented in an accessible, non-technical manner. Furthermore, stories may resonate more with policy-makers, people involved with human services, or with tribal and other communities whose ways of knowing rest more in stories than in tables of regression coefficients (Patton, 2002; Tribal Evaluation Workgroup, 2013). Well-chosen and thoughtfully presented direct quotes by program participants can enliven an otherwise technically dense report, and peoples' stories often are compelling and can make the report more relevant for a wide range of audiences.

TEXT BOX 1: The Role of Qualitative Methods in Program Evaluation

- Theory of Change/Logic Model Development
  - Logic model development: Generating *detailed information* about program inputs, activities, and outcomes
  - Theory of change: Explaining *how* and *why* a program exerts its effects
- When Established Measures Are Inappropriate or Do Not Exist
  - In early stages of research when unclear what measures are best
  - With populations for whom valid standardized measures do not exist
  - With populations who may be more comfortable with story-telling and narrative
- Studying Program Implementation: Gathering *detailed information* about:
  - How well defined practices and programs are
  - Level of buy-in/readiness for the program among community members/stakeholders
  - Program provider's staff hiring, training, coaching, and evaluation practices
  - Efforts made re: organizational change necessary to implement the program
- Opening the "Black Box" of Program Effects: Unpacking *why* the program had the effect it did by exploring:
  - Experiences of clients and staff
  - Contextual changes that might affect outcomes
  - What outcomes mean to program participants
  - Unintended programmatic side effects
  - How well/faithfully the program was implemented
- Making Research Reports More Accessible

- Present some findings in a non-technical format (i.e., quotes, not statistics)
- Stories and narrative about participants' experiences resonate with policymakers and other stakeholders
- Quotes and stories enliven technically dense reports and make them more readable

---

Summary

The starting point for deciding whether or not to invest in qualitative research is *always* the research questions to be answered and consideration of the type of data needed to answer them. The preceding section on the role of qualitative methods in program evaluation provides broad guidance on this point. If the research questions require complex or detailed information that measures relying on short-answer, quantifiable data cannot capture — if, in other words, what is needed is not simply the numbers, but the stories and explanations behind them — the project probably needs to collect qualitative data. If the research questions are exploratory in nature — that is, there is not yet enough clarity about what the study will find to select closed-ended measures — the project probably needs to collect qualitative data. It is important to stress, however, that employing qualitative methods in an exploratory or early-stage study does not obviate the need to justify the choice of research questions and methods. The findings from prior, related research should be used to make the case for the research questions to be answered next and the need for qualitative (or quantitative) methods.

No one evaluation method will answer every question, and a comprehensive program of evaluation likely will occur over multiple studies as a program is developed, implemented, and replicated (Framework Workgroup, 2014; National Research Council and Institute of Medicine, 2002). Both quantitative and qualitative methods have roles to play and can make different contributions to a program of evaluation. Philosophical arguments favoring one set of methods over another have long since been resolved by pragmatic considerations: Which method is used depends on what questions are being asked: on what it is, exactly, that we want to know (Miles & Huberman, 1994; National Research Council and Institute of Medicine, 2002; Patton, 2002; Sechrest & Sidani, 1995; Shadish, 1995b). It is not necessary that every program evaluation include both qualitative and quantitative methods; sometimes just one of these is appropriate for the questions to be answered at a given stage of the program's development. What is necessary, however, is that whatever method of inquiry is used, it be implemented as rigorously, credibly, and transparently as possible. The following sections of this document were created to help you work with evaluators to get the most out of a program evaluation that includes qualitative methods.

.

PART TWO: GETTING STARTED

After carefully considering what questions need to be answered by an upcoming round of funding for program evaluation, it has been determined that qualitative methods should be part of the funded project/s. If you will be involved in developing the funding opportunity announcement (FOA)or contract request for proposals and/or participating in the review of applications or development of the criteria used to rate them, you will benefit from reading all the rest of this document.

But let us start with some early tasks first: considering the time and budget needed for the funded projects and choosing an evaluator with the necessary experience in using qualitative methods. Each of these topics will bear on how you develop and word the funding announcement and select the winning applicant.

Researcher Experience

The importance of hiring an evaluator with solid experience in qualitative methods cannot be overstated. As Michael Quinn Patton reminds us: "For better or worse, the trustworthiness of the data is tied directly to the trustworthiness of the person who collects and analyzes the data — and his or her demonstrated competence" (2002, p. 570). Certainly, the importance of having only skilled and experienced professionals conduct evaluation research involving any kind of data — qualitative or quantitative — is critical if the data collected are to be credible and trustworthy, but qualitative methods are particularly dependent on the skill of the researcher (American Evaluation Association, 2013; Patton, 2002; Sofaer, 2002). Therefore, it is critical that an evaluation proposing to incorporate qualitative methods be led by a researcher who has the necessary competence to implement these approaches well.

Prior experience using qualitative methods and published research that includes the use of these methods are some things to look for, and the biographies and resumes of the principal investigator and key members of the research team should reflect the relevant credentials and prior experience (Morse, 2003). If the project proposes to use graduate students or other junior staff for the field work, it is essential that these individuals be trained and their work supervised by experienced researchers whose time commitment to the project is sufficient to provide such training and supervision (Morse, 2003; Patton, 2002; Silverman, 1990). The work plan or evaluation proposal should indicate how this training will be accomplished, who will deliver it, and how much training (and in what methods) will be received. Some experts suggest that an interdisciplinary research team will ensure a diversity of perspectives and work to reduce the possibility of bias (Cohen & Crabtree, 2006; Silverman, 1990), although this approach may not always be feasible. Another approach to research team composition is to include both those who are members of the disciplinary or professional group being studied (i.e., "insiders") and those from other fields ("outsiders"). Insiders may more easily gain access to the site, have more credibility with informants, or be better able to distinguish unusual events from typical practice. However, outsiders may be better able to maintain objectivity or discern unspoken, shared assumptions among program staff. Therefore, some experts recommend an evaluation team that includes both insiders and outsiders, if possible (Cohen & Crabtree, 2006; Morse, 2003).

Questions to ask about researcher credibility

1. What experience does the evaluator have with using qualitative methods for a program evaluation? In which specific methods (i.e., focus groups, interviews, observation, document review) does the evaluator have expertise? How well does this expertise fit with the design of the evaluation?

2. Given the goals of the qualitative aspects of the evaluation, does the research team reflect an appropriate mix of disciplinary backgrounds and members with insider or outsider status to accomplish the data collection and analysis?

3. If graduate students or junior staff will be used for field work (going on site to observe, conducting interviews, etc.), what are the plans for their training and supervision? Does the project budget include an adequate time commitment from the principal investigator or other senior research staff for these activities?

Budgeting and Time Planning

The labor- and time-intensiveness of qualitative data collection and analysis require that the project budget allows for such costs and that project timelines are adequate. Qualitative research can be expensive, and the major item will be personnel costs (Morse, 2003; Office of Behavioral and Social Sciences Research, 1999). Furthermore, compared with quantitative research, qualitative projects typically require less time for instrument development, but *more time* for data analysis (Miles & Huberman, 1994; National Science Foundation, 1997). Preparation of verbatim interview or focus group transcripts and narrative observational field notes also takes time (Office of Behavioral and Social Sciences Research, 1999). It is essential that the project budget not underestimate analysis time for qualitative data.[4]

Although each project differs somewhat, Miles and Huberman (1994) draw on long experience to suggest the following broad time-planning guidelines:

- For each hour of audio (interviews, focus groups), expect to spend 4 to 8 hours on transcription, depending on the level of detail in the interview and how familiar the transcriber is with the content.
- For each day of observation, plan on 2-3 days to prepare field notes for analysis.
- Coding: Multiply the time spent collecting the data (i.e., the length of the interview) by a factor of 1 or 2. In other words, it will take at least 1 hour, and possibly as long as 2 hours, to code a one-hour interview, depending on the complexity of the coding scheme. It will take 1 or 2 days to code a single day's worth of observational data, etc.
    - Bear in mind that a one-hour interview will, on average, result in 10-15 single-spaced pages of text (Patton, 2002). If the project plans to interview 30 people, that's 300-450 pages of data.
- Preparing data displays: Again, multiply the time spent collecting the data by a factor of 1 or 2 to prepare data displays and complete other analytic tasks.

[4] Part Three of this document provides more information on the elements of qualitative research methods, including data collection and analysis. Reading that section will help you understand some of the procedures referred to in the time and budget planning discussion.

Miles and Huberman caution that the above estimates apply to single cases (for example, a single program site). If multiple program sites are involved, be sure to consider each day or hour in the field *at each site* in planning costs.[5] Also, assume evaluators will need additional time for cross-site analyses. Morse (2003) suggests having a "contingency allowance" built into the budget in anticipation of interviews that are lost or unusable because respondents were not sufficiently informative or because recording equipment failed, as well as to allow evaluators to explore fruitful avenues that may emerge during fieldwork.

Additional direct costs to consider include:[6]

- Staff travel to conduct on-site field work
- Audio (and possibly, video) recording equipment
- Qualitative analysis software
- Staff time for preliminary coding work to establish inter-rater reliability and to allow for revision of the coding scheme
- Thank-you gifts for interview and focus group participants.

University-based researchers may be able to draw on graduate students to carry out much of the field work, and this approach can be cost effective. However, Miles and Huberman (and personal experience) caution against a too-strict division of labor between junior and senior staff, noting that it is difficult to fully understand qualitative data during analysis and interpretation for someone who has not spent any time in the field (1994). Be sure the budget allows for the time and attendant labor costs of senior staff.[7]

Finally, time plans "usually suffer from excessive optimism," say Miles and Huberman (1994, p. 47). If project timelines are too short, evaluators may be tempted to formulate conclusions before adequate analysis and interpretation have taken place (Cohen & Crabtree, 2006). One strategy that can help avoid this problem is to have evaluators do the math on time (and related cost) planning before going into the field, but then review it after the first round of field work has been completed to make sure the initial estimates are on track.[8] If they appear overly optimistic, the evaluators should consider reducing the amount of data collected, rather than giving short shrift to the analysis (Miles & Huberman, 1994).

[5] No doubt there will be differences of opinion regarding these estimates, and differing levels of complexity in among projects — as well as the skill and experience of the research team — will affect the amount of staff time needed. Unfortunately, there are few concrete guidelines available for estimating staff time needs in qualitative research; thus, what are offered above should be regarded as very general estimates.

[6] Much of this list is taken from Morse, 2003; some is from personal experience.

[7] Although Miles and Huberman do not address the matter of data collection by program staff, it is likely the same principle applies: The individual, presumably an experienced consultant, who does the data analysis and write-up should spend at least some time personally in the field.

[8] The "first round" might consist of the field work at the first of several program sites, for example, or after a small portion of planned interviews/focus groups have been conducted.

Questions to ask about time and cost estimates[9]

1. Do the estimates of on-site time and attendant costs appear reasonable to accomplish the planned data collection with the number (and expertise) of staff who will be doing the field work? Consider:
    a. If estimated travel costs accurately reflect the scope of planned field work.
    b. Who will be doing the field work (i.e., how many junior and senior staff) and how well the budget accounts for the personnel costs, including those of senior staff.

2. Do the estimates for data preparation and analysis appear reasonable? Consider:
    a. Given the number of interview subjects and/or focus groups planned, if the time and cost estimates for transcription appear reasonable.
    b. Who will be doing the transcription: If someone other than the researchers (i.e., someone unfamiliar with the content of the interviews or focus groups), be aware that transcription time will run longer.
    c. If time estimates for data coding and display seem adequate given the volume of data to be collected and the complexity of the planned coding.
    d. If analytic steps in addition to basic coding will be taken and how well the work plan and budget account for these.
3. Does the budget allow adequately for other direct costs such as qualitative analysis software, audio/video recording equipment, and respondent thank-you gifts?
4. Do the evaluators have a fallback plan — or are they willing to consider one — if the initial time and costs estimates prove unrealistic? What specific changes would be made to bring the project in on time and within budget?[10]

If the research budget is tight, consider a small, but very well-targeted inquiry. Using program staff to collect data — with expert consultation and oversight — can be a cost-effective option (see the section on using program staff for data collection in Part Three of this document.) Also consider performing the analysis on summary documents, rather than verbatim interview transcripts (see the section on data analysis in Part Three of this document). None of these options is ideal, but if the research questions demand qualitative data, they may be the best that can be done, at least as a first step. Implemented thoughtfully, consistently, and transparently, even small efforts can yield useful insights that will inform subsequent, perhaps more extensive, studies of the program.

Summary

Engaging an experienced evaluator and ensuring that the funding, and the selected project's budget and timeframes, are adequate to support a good quality project are critical first steps that will have a direct bearing on the final credibility and validity of the study's findings. Of course, another critical element is that the research methods are implemented as rigorously as possible. The next section of this document presents a high-level overview of qualitative research methods.

---

[9] Part Two of this document, which provides a basic primer on qualitative methods may be useful in clarifying some of the terms used here and in providing an overview of the steps involved in qualitative data collection and analysis.

[10] It is possible that the evaluator would be able to be specific about changes only after work has begun and the nature of the challenges is more clear. Early on, it may make sense simply to establish the evaluator's flexibility.

PART THREE: A BRIEF PRIMER ON QUALITATIVE METHODS

Before presenting an overview of qualitative research methods and how to assess their quality, a couple of clarifications may be helpful. First, there are competing perspectives on what constitutes good qualitative research; that is, the criteria for judging these methods differs among qualitative researchers. The questions and criteria presented here, although driven more by practicality than philosophical purity, resemble most closely what Patton (2002) calls "traditional scientific research criteria."[11] Used for evaluation, this approach works to describe programs and participants' experiences as accurately and objectively as possible. Researchers in this tradition will use concepts and terminology — hypothesis-testing, variables, etc. — that align well with quantitative research methods. This approach also represents the most typical framework for government-funded research (Patton, 2002) and, therefore, may be most useful for federal staff tasked with overseeing funded evaluation projects.

Second, good qualitative research is just as rigorous as quantitative, but the criteria for judging rigor are less standardized than they are for quantitative studies. There are, for example, no mathematical procedures for determining the correct sample size or level of confidence in conclusions, no hard and fast rules for which analytic method to use depending on the distribution or characteristics of the data (Miles and Huberman, 1994; National Research Council and Institute of Medicine, 2002; Patton, 2002). That said, there are ways to determine the credibility and quality of qualitative research. It is not true that "anything goes," or that qualitative research is simply a matter of talking with people and presenting a hodge-podge of quotations. A high quality study will include thoughtful and consistent data collection and analytic procedures that are transparent and verifiable (James Bell & Associates, 2009; Sofaer, 2002).[12] What this document will try to do is provide some information on what well-implemented qualitative methods consist of and to suggest questions you can ask grantees and their evaluators.[13]

Research Design: The Structure of the Study

Let us begin by distinguishing two separate concepts - research design and type of data. The terms "qualitative" (and "quantitative") describe *types of data* or the methods used to collect and analyze those data, whereas the design of a study is the overall architecture within which

[11] For those interested in reading about other philosophical perspectives, see Patton (2002), pages 542-553 and the Robert Wood Johnson qualitative research website at: http://www.qualres.org/HomeEval-3664.html.

[12] NIH's Office of Behavioral and Social Sciences Research suggests asking evaluators to do more than simply name various components of their study (such as "purposive sample," "semi-structured interviews," and the like) but to include also a clear, jargon-free description of what they plan to implement. See http://obssr.od.nih.gov/pdf/qualitative.pdf.

[13] The focus of this methods discussion is on qualitative methods. Certainly, there are some research tasks and considerations that apply to all studies, whether they rely on qualitative methods, quantitative methods, or both. Such matters as reviewing existing knowledge to lay the groundwork for the current study or carefully formulating and bounding the research questions apply to program evaluation in general. The Children's Bureau has prepared a useful framework for thinking about the applicable evaluation questions at all stages of a project. See Framework Workgroup, 2014.

data collection and analysis are carried out (James Bell Associates, 2009; OPRE, 2010; Secrest & Sidani, 1995).

When we speak of randomized control trials (RCTs), quasi-experimental designs, interrupted time series, case studies, and the like, we are talking about research design. Qualitative data may be collected within any of these research designs, including RCTs and quasi-experimental designs (National Research Council and Institute of Medicine, 2002; Patton, 2002). The choice of research design should be driven by the research questions and goals, and this holds true regardless of whether qualitative or quantitative data (or both) are to be collected.

However, as a practical matter, program evaluations use qualitative data most often (but not exclusively) to describe or document aspects of the program (i.e., in formative evaluations), rather than to demonstrate causation (i.e., program effectiveness or outcomes). Therefore the design of qualitative data collection quite often will be non-experimental or "naturalistic."[14] For example, even if program outcomes are being determined using an RCT or quasi-experimental design, the implementation or process component of the study may be designed as a case study. Sometimes, you may see qualitative data included in an outcomes evaluation that uses a control or comparison group, perhaps to gather and compare more detailed data from a subset of program participants.

Questions to ask about research design

The important questions to ask about research design revolve around whether or not the design supports the goals of the study.

1. What is the goal of the qualitative component of the study; that is what information is it intended to provide? How well is the goal supported by a review of prior, relevant research?
2. Have the evaluators stated clearly how the design approach will provide the desired information?
3. If the qualitative data are being collected under an RCT or quasi-experimental design, will the qualitative data be collected from both experimental and control/comparison groups?
    a. If not, why not? What interpretive or explanatory benefits will be lost by not collecting these data on both groups?
    b. What are the cost/time trade-offs with a more extensive qualitative data collection?

There is not one correct answer to these questions; the goals of the study (as well as time and budget constraints) should determine the best approach. However, if the qualitative data are intended to provide interpretational context for quantitative findings exploring differences between experimental and comparison/control groups, one would expect to see the qualitative measures collected on both groups — or a subset of each group — in order to support the

[14] Although demonstrating causation in a program evaluation - i.e., determining whether or not the intervention "caused" the observed outcomes — is most typically accomplished with quantitative data and research designs such as RCT or quasi-experiments — there is an analytic technique called “causal modeling” that works to determine cause-effect relationships using qualitative data (Miles & Huberman, 1994). This approach does not yield quantitative estimates of the strength of the relationship, and the analysis process is labor-intensive. It is not often used to demonstrate program effectiveness, and its use is beyond the scope of this document.

comparison. But if, for example, the qualitative findings are intended primarily to capture the perspectives of program participants, then omitting similar measures for the comparison/control groups may make sense.

Sampling: What Gets Measured

The logic of sampling for a qualitative inquiry is entirely different from that used for quantitative measures.[15] Samples drawn to support quantitative data collection and analysis are designed to maximize statistical probability and support generalizing from the sample to a population — and being able to quantify the "confidence" researchers have in those conclusions. For this reason, some version of random sampling, to minimize selection bias, and the creation of comparatively large samples, to enhance statistical power, are desirable for quantitative analysis.

But qualitative analysis focuses on developing in-depth information on and insight into a limited number of cases. Therefore sampling is carried out to include what Patton terms "information-rich cases" and typically is not done randomly, but *purposefully*.[16] A full discussion of the various sampling approaches used to collect qualitative data (there are over a dozen) is beyond the scope of this document. Suffice it to say, qualitative samples tend to be comparatively small, are usually (but not always) non-random, and do not support sample-to-population generalizability. These characteristics do not signify a lack of scientific rigor so long as the specific sampling approach supports the goals of the study. Text Box 1 describes a few types of purposeful sampling strategies.

Sampling Strategies in Qualitative Research

TEXT BOX 2: EXAMPLES OF PURPOSEFUL SAMPLING STRATEGIES

- *Heterogeneity samples* (sometimes called "maximum variation" samples) work to ensure representativeness across sources of variation thought to be important such as urban/rural programs or participants, age groups, gender, community characteristics, program's stage of development, etc.
- *Homogeneous samples* are often used for focus groups to generate in-depth information about a subgroup of participants who share key characteristics that allow them to reflect together on specific issues.
- *Extreme or deviant case sampling* may be used to explore unusual cases, perhaps contrasting program clients who had extremely good outcomes with those who had extremely negative ones, in order to tease out an explanation of what drove these very different outcomes.

---

[15] Most of the discussion of sampling here is taken from Patton (2002), pp. 230-246) and from Miles and Huberman (1994), pp. 27-34. Other citations will be noted as used.

[16] "Purposeful" is the term Patton prefers. Miles and Huberman use "purposive" to describe the same idea. You may also see the terms "judgment sampling," "theoretical sampling," or "focused sampling." All these terms go to the same general idea: that cases to be studied are selected not randomly, but with some goal or purpose in mind.

- *Snowball or chain sampling* is best used to locate key informants or other information-rich cases. The researcher will ask each informant to suggest others who are particularly well-positioned to speak on the program issue under study. This approach may also be used to locate politically important cases or informants, or informants who may be difficult to locate or identify, such as homeless or isolated individuals.

Sources: Cohen & Crabtree, 2006; Miles & Huberman, 1994; Patton, 2002; Silverman, 1990.

The list of sampling strategies in the text box is by no means exhaustive; rather it is intended to suggest the logic of purposeful sampling and the range of study goals it may serve.[17] However, there are two other sampling approaches encountered in program evaluation plans and proposals that do not resemble any of the above: convenience sampling and random purposeful sampling.

- *Convenience sampling* selects cases based on the ease and speed with which they can be located. This type of sampling is used quite often because it is easy and inexpensive, but it is probably the least desirable sampling approach because it may unknowingly omit the most important cases. Convenience sampling is not considered purposeful sampling (Cohen & Crabtree, 2006; Patton, 2002; Silverman, 1990).
- *Purposeful random sampling:* Where a program evaluation is using both qualitative and quantitative methods, you may see a sampling approach that selects a large random sample for the quantitative outcomes measures, but also includes qualitative measures on a small subset of participants who are chosen randomly from the larger sample. Randomly selecting a few cases for in-depth analysis or to provide concrete examples of participants' program experiences lends credibility to the findings, although it does not permit generalization to the entire client or staff population.[18]

However the sampling is to be done, it should be consistent with the purpose of the study, be specified in advance (i.e., at the proposal or research plan stage), and include a statement about the degree to which findings can be generalized (Silverman, 1990).

Sample Size

Unlike in quantitative research, where there is a formula for determining the sample size needed for statistical power at a given level of confidence, there is no formula for determining the "correct" size of a qualitative sample. Sample size in qualitative research is not judged by the same criteria as it is in quantitative research because statistical power is not the goal (Patton, 2002). Instead, the researcher needs to consider the trade-offs between breadth and depth, and that can be done only by considering the purposes of the qualitative portion of the study. If the goal is to capture variation — i.e., breadth — across cases (programs, individuals), that may call for a larger sample than one intended to explore a narrow phenomenon in depth. The sample

[17] For an overview of a wide range of sampling strategies and when each is best used, see the discussion at http://www.qualres.org/HomeSamp-3702.html.

[18] Patton refers to this approach as "purposeful random sampling." The term matters less than the idea that qualitative sampling sometimes may be done randomly, but that the typically small sizes of qualitative samples still limit sample-to-population generalizability.

size should provide what Patton terms "reasonable coverage" of whatever is to be studied, and it can be a good practice to ask evaluators to specify up front the *minimum sample* to be studied. This number can be increased if "saturation" is not reached once data from the minimum sample has been collected. (Sample "saturation" is reached when data collected from additional subjects no longer contribute any new information.) It is important to bear in mind, however, that as the sample size increases, as a practical matter, the data one can collect become thinner. Miles and Huberman remind us, if a qualitative sample gets so large that the data have to be simplified to the point where they could be collected via a survey, why not just do the survey (1994)?[19]

Questions to ask about sampling:

1. Can the evaluator explain how the sampling strategy meets the goals of the study? Why were other strategies ruled out as less desirable?

2. Does the sampling strategy make intuitive sense? That is, does it seem likely to include the most information-rich cases or informants and do so in an organized, systematic manner? For example, in an outcomes study, the most information-rich cases likely would be a thoughtfully-chosen sample of program recipients; an implementation study would include key actors in implementing the program, etc. Have any potentially important (i.e., informative) individuals and groups been overlooked by the sampling strategy?

3. If the plan or proposal calls for a "convenience sample," can the evaluator explain why a more systematic approach to sampling was not adopted? If resources are limited, it might be that a smaller, but purposefully identified, sample would provide more credible findings than a larger convenience sample.

4. Has the evaluator provided a minimum sample size for the qualitative data collection? Can the evaluator explain how this size of sample will provide reasonable coverage, achieve representation of the individuals/cases to be studied, or achieve saturation?

Qualitative Data Collection: Methods and Other Considerations

This section of the document takes up the collection of qualitative data, both the methods by which it is conducted and a couple of additional considerations that will bear on the quality and credibility of the data. This information will be useful when reviewing grant applications and evaluation plans insofar as it will familiarize you with just how qualitative researchers go about gathering their data. The discussion begins by outlining the methods by which qualitative data are collected. Next, we consider the problem of researcher bias that can arise with on-site field work. Finally, the matter of using program staff to collect data is considered.

[19] As Morse (2003) points out, good qualitative research usually is more expensive than quantitative research. Therefore, it makes sense to consider carefully what qualitative methods will contribute to the project and use them when their unique strengths are called for.

Qualitative evaluation data have three sources: interviews (including focus groups), observation, and document review. Because the most frequently encountered qualitative data within program evaluations are interviews and focus groups, this section will spend the most time considering them, touching on the other two only briefly. However, a particularly strong approach to any study is to triangulate data collection methods and data sources; that is, to combine multiple approaches to and sources of data within a study. Doing so, if the project budget permits, can increase the validity of the data.[20]

Interview techniques are appropriate when we want to know about things that cannot be directly observed (Patton, 2002, p. 340). Typically, interviews are used to tap the experiences and perspectives of program participants, staff, community members, or other individuals involved in some way with the program. Exactly who is interviewed or included in a focus group is a sampling issue (see preceding section). In this section, we discuss different interview approaches and provide some guidance on how to tell if these methods are being implemented appropriately.

Interviews and Focus Groups

Text Box 2 lists the four main types of interviews. As a practical matter, many interviews will be a combination of these techniques. Patton notes, for example, that an interview may begin with a structured protocol, but allow the interviewer to spontaneously pursue other topics that are relevant toward the end (2002).

TEXT BOX 3: TYPES OF INTERVIEWS

- Unstructured or informal interviews (sometimes called ethnographic interviewing): No question list prepared in advance; instead questions arise spontaneously, perhaps in response to something noted in observation or in a document.
    - Strengths: affords maximum flexibility and responsiveness to context, therefore useful in early stages for developing ideas or when it is not clear what is important to ask about; helps build rapport with informants.
    - Weakness: time-intensive as more than one conversation with any given informant may be needed; also may be more difficult to synthesize and analyze.
- Semi-structured interviews: Use an interview guide which includes questions or issues to be asked about, but does not specify exact wording. Interview guides may also include optional "probe" questions to remind the interviewer to follow up on specifics as needed.
    - Strengths: allows interviews to be focused, while still allowing for respondents' perspectives to emerge or other, relevant issues to be explored; useful if time

[20] Denzin (1978) identifies four types of triangulation: by method, by data type, by investigator, and by theory. Triangulation of data and method are what is being referred to above. By "method," we are talking about different data collection methods (interviews, observation, document review, etc.). Triangulation of data involves drawing on different data sources; i.e., talking with or consulting a range of informants, sites, or documents. These two types of triangulation also could include using both quantitative and qualitative methods and gathering both types of data. The topic of triangulation is taken up at greater length in Part Four of this document.

allows only one interview per respondent; interview guides can be prepared with more or less detail, as circumstances warrant.
    - Weaknesses: Methodologically strong and often used in research, but requires an experienced interviewer to know when/how to probe, so that critical areas are adequately covered and potentially important, but unanticipated, ideas are pursued.
- Structured or standardized, open-ended interviews: Use an interview protocol with questions worded in advance and asked the same way across respondents.
    - Strengths: Easier to compile and analyze data; minimizes variation among interviewers; most time-efficient; may be more acceptable to institutional review boards (IRBs) or other approving bodies because exact questions are specified in advance.
    - Weaknesses: Interviewers cannot pursue unanticipated, but possibly important topics that arise during the interview; less ability to explore respondents' individual circumstances; requires that evaluators know in advance what will be important to ask about.
- Focus groups: Typically include 6-10 persons discussing a specific topic for 1-2 hours. Although it is not always practical, some experts suggest that participants be strangers to one another so that existing personal relationships do not inhibit candor. An interview guide, similar to what is used with semi-structured individual interviews, generally is used. Ideally, at least two skilled individuals (and some experts recommend three or more) will run the group, with one asking the questions and the others taking notes, dealing with audio recording, etc.
    - Strengths: Cost effective because several people are interviewed in the time it would otherwise take to interview one person; participant interactions can improve data quality by providing a check on extreme views; facilitates gauging the degree of consensus or variation in viewpoints.
    - Weaknesses: The number of questions and range of issues that can be explored is very limited; individual respondents' response time is limited to allow for others to participate; not suitable for controversial or very personal issues; requires highly skilled facilitators.

Sources: Patton, 2002; 342-348, 385-390; Cohen & Crabtree, 2006 at http://www.qualres.org/HomeInte-3595.html.

Asking Open-Ended Questions

Regardless of the type of interviews to be conducted, qualitative data collection via interview relies on *open-ended questions*. The types of forced choice questions (yes/no, multiple choice, Likert scale) often found on surveys, whose responses lend themselves well to quantitative analysis, are not the mainstay of qualitative interviews.[21] That said, some surveys

[21] The term "interview" may be used by survey researchers who contact respondents in person or by phone. However, surveys generally rely primarily on closed-ended questions. It is important to distinguish between a survey interview and true qualitative interviewing.

may include open-ended questions, and these may be analyzed using qualitative analysis techniques.

Open-ended questions should be worded to allow respondents to use their own words in responding. That means the wording should not suggest categories of response (more, less, often, rarely) or be framed as Yes/No questions (Patton, 2002). Most often, questions will open with certain words — who, what, where, when, how.[22] Good interview questions are clear in their intent, ask about only one issue at a time, and are neutrally worded (Patton, 2002).

If the evaluators plan to use an interview guide (semi-structured interviews or focus groups) or a standardized interview protocol, it will be prepared in advance of the fieldwork, and you can ask to see it. Indeed, if the evaluation is conducted under contract to the federal government, you must review instruments such as interview guides and comply with Paperwork Reduction Act requirements for public notice and OMB review.[23] If the questions are poorly worded or if they do not seem to require qualitative techniques (i.e., they don't require asking open-ended questions), you may wish to discuss the methods choices with the evaluators.

Capturing Interview Data

It is highly advisable to audio-record interviews and focus groups to capture exactly what informants say as well as to allow the interviewer to pay close attention to responses and formulate follow-up or clarifying questions (Patton, 2002). Certainly, if verbatim transcripts of interviews and focus groups will be prepared, it is essential that these be audio-recorded.

A lower-cost alternative to preparing full transcripts is to have interviewers use a data capture form or template that summarizes respondents' replies, capturing a limited number of important quotes verbatim. In such cases, interviewers (or focus group note-takers) would use the notes they wrote during the interview or group to complete the data capture forms. However, even if verbatim transcripts will not be prepared for analysis, the audio recordings are strongly advised because they will be needed to perform quality and validity checks and again at the coding stage to check inter-rater reliability. Therefore, project budgets should allow for audio equipment costs.[24] Even when audio recording is used, "strategic and focused" note-taking during the interview is recommended to provide an alternate data capture system in case audio equipment fails, to capture key ideas, as well as helping to pace the interview (Patton, 2002).

Qualitative Observation

Observation is another data collection technique that can produce qualitative data in the form of detailed field notes.[25] Observational data can tap into phenomena that program

---

[22] Patton (2002) and Becker (1998) caution against asking "why" questions for a number of reasons, one of which is that it may imply some disapproval or judgment about a response. Generally, such questions can be reworded to get at what the evaluator is interested in. For example, if we want to know the reasons a participant joined a program, the question can be asked as, "Tell me what about the program led you to join" or some similar wording that poses less risk of putting the respondent on the defensive.

[23] See Office of Management and Budget, Office of Information and Regulatory Affairs website: https://www.whitehouse.gov/omb/oira

[24] The IRB may also require that separate informed consent be given expressly for recording. This consent generally is sought after the subject agrees to be interviewed.

[25] Qualitative observation differs from observational techniques used to complete a measurement instrument or tool that will be scored and analyzed quantitatively. That is, it is not observation itself that is qualitative; rather the form of the resulting data - numerical scores or field notes in words — that determines if it is quantitative or qualitative.

participants may not express in interviews — or even be fully aware of (Patton, 2002). Spending time observing program activities (services, meetings, etc.) or having informal conversations with key informants can help an evaluator gain a clearer sense of context, and that information can aid in interpreting what is learned in interviews. Observation may be particularly valuable in assessing program implementation or capturing interpersonal dynamics among participants.

The downside of observation is that it is exceptionally time- and labor-intensive: Research staff must spend time in the program, and the preparation of field notes capturing what is observed or said (i.e., direct quotes) can take even longer than the time spent on site (Miles & Huberman, 1994; Patton, 2002). Audio or video recording may be used instead of or in addition to written note-taking, but some transcript or narrative describing what the recordings contain will need to be prepared to allow for analysis (Cohen & Crabtree, 2006; Patton, 2002). Therefore, if this technique is proposed, the evaluation budget should reflect the necessary time and costs.

Document Review

Reviewing documents is another qualitative research strategy that offers a number of advantages to the evaluator. Documents can convey a sense of events that began prior to the evaluation, documenting early program development and implementation stages and the decisions that were made along the way. Insights may be gained comparing public documents such as reports with private memos or with what participants report in interviews.[26] Documents also can help to stimulate questions that may be pursued later through interviews and observation (Patton, 2002). When your evaluators negotiate access to program sites, they should include access to documents that can inform the evaluation.

These advantages aside, procedures for implementing document review are less well developed than for other types of qualitative data collection (Cohen & Crabtree, 2006). Miles and Huberman (1994) recommend the use of a document summary form to capture key information: After being filled out by the researcher, the form is attached to the document. Key information captured will vary from project to project, but may include the document name, the date and place it was acquired, the document's significance, and a brief summary of its contents. The researcher's comments or reflections on the document may also be captured. These forms can later be coded for analysis.

On-Site Data Collection and Researcher Bias

As you probably have inferred by now, qualitative research relies only minimally on standardized instrumentation.[27] Instead, data are collected by a researcher or team of researchers who come into direct contact with the individuals and settings under study, and

[26] The term "documents" includes a broad range of items including reports, memos, client records, email. meeting minutes, program brochures, organizational rules and policies, budget and financial records, etc. Access to and use of certain non-public information may be subject to informed consent procedures and confidentiality restrictions.

[27] The term "standardized" instruments is being used to refer to tests and other measurements that have been designed and are used in many different studies and settings, and whose validity and reliability have been tested and quantified. These measures ask all respondents to answer the same set of questions, and typically those responses are scored in the same manner across all respondents. Standardized educational tests are one example. Certainly, within a given study using qualitative methods, there may be small-scale standardization in that the same questions may be asked of all interviewees, but the questions most typically are formulated for that study specifically, and the analysis will reflect the particular concerns and questions that drive the project.

typically do so for a more prolonged period than do quantitative researchers. Therefore, in some sense, the researchers are the measurement instrument, and for that reason, their experience and freedom from bias are critical elements in the credibility of the data that they collect (Miles & Huberman, 1994; Patton, 2002).

The possibility of researcher bias arises especially when research is conducted in the field, and the researcher is interacting with those being studied. This problem is not exclusive to qualitative methods, but because qualitative research typically is conducted in field settings — program sites, communities, etc. — and involves more extended interaction with informants and subjects than do quantitative research methods, the possibility of researcher bias is heightened. This potential for bias takes two main forms: The effect of the researcher on the site/informants being studied, and the effect being on site may have on the researcher (Miles & Huberman, 1994; Patton, 2002; Silverman et al., 1990).

Program evaluations may be particularly vulnerable to bias stemming from the researcher's presence at the site because program staff and participants being studied have a stake in what the evaluator may find (Patton, 2002). They may respond by being more attentive to participants or procedures than usual or they may become anxious about the evaluation and slightly alter their behavior and practices. Although experts caution researchers not to overestimate these effects, they do recommend some strategies for minimizing the possibility of this type of bias, including making sure program staff are clear on the intentions of the research, what will be studied, how information will be collected, and what will be done with the information. Evaluators may also consider conducting some or all of the interviews and focus groups off-site (Miles & Huberman, 1994; Patton, 2002). One other approach is for researchers to spend as much time as possible at the site in order to better fit in and allow program staff to become more comfortable with the evaluation staff.

However, the increased costs of more time in the field may make this approach infeasible. In addition, spending more time on site increases the likelihood of the second form of bias — that the researcher's objectivity may be compromised (Miles & Huberman, 1994; Patton, 2002). Over longer periods of contact, the researcher may also develop relationships with program staff and wish to avoid findings that open them to criticism (Becker, 1970). Strategies evaluators can take to minimize the likelihood of fieldwork interfering with their independence and objectivity include being sure to interact with a wide range of participants and stakeholders (to minimize being unduly influenced by any one group or perspective, particularly program leadership or other "elite" participants), being sure to observe/report on a wide variety of events and circumstances, sharing interview and field notes with others on the evaluation team, and reflecting on/recording noted changes in perspective as the research proceeds (Becker, 1970; Miles & Huberman, 1994; Patton, 2002). And, although it directly contradicts the advice for minimizing the researcher's effect on the site, Miles and Huberman (1994) also advise researchers to spend time away from the site.[28]

The main recommendation for both types of researcher bias is for evaluators to be aware of this potential and make their concerns explicit. Strategies for accomplishing this, in addition to those mentioned above, include holding early and frequent team meetings to identify and

[28] Although the advice about spending time on site can be contradictory, as a practical matter, the decision may solve itself: Most evaluation budgets will allow for just enough on-site time to accomplish the planned data collection, no more and no less. However, where on-site time is very short or very long (or where there is an existing relationship between the program and the evaluator), it is important to be aware of the potential for researcher bias in either direction and ask evaluators how they plan to address this possibility.

discuss assumptions and biases. Researchers may also wish to have reports read and reviewed by informants and other key stakeholders (Silverman et al., 1990). Such review may potentially introduce the biases of these reviewers, but may tend to offset potential researcher bias. Finally, evaluators must take responsibility for thinking about potential bias and devising approaches to minimize it (Patton, 2002).

A Few Words About Data Collection By Program Staff [29]

Ideally, qualitative data will be collected by experienced individuals who do not work for the program under study. However, you may encounter proposals that rely on program staff to collect data, and it is important to understand the advantages and disadvantages of this approach.

Advantages:

- Less costly than data collection by external groups/individuals.
- Program staff may be more invested in — and less resistant to — the evaluation.
- Opportunity for program staff to reflect upon and understand their program.
- Staff knowledge of and rapport with clients or community members may enhance the utility and value of the data collected.

Disadvantages:

- If program clients are interview/focus groups subject, they may be less candid in their responses.
- Diminished confidentiality for clients/subjects.
- Staff interests in seeing specific outcomes may bias the data.
- Methodological rigor and credibility of findings may be questioned by external audiences.

If available resources or other pragmatic considerations necessitate using program staff to collect data, the evaluator's role in data collection may consist primarily of consultation to that component of the project.[30] In this capacity, the evaluator can and should take several steps to maximize the credibility and consistency of data collection, including:

- Designing a structured interview/focus group protocol.
- Developing a standardized data capture template for document review, and, if verbatim transcripts are not utilized, also for interviews and focus groups.
- Training data collection staff in the use of the protocols and templates.
- Arranging data collection so that program staff, insofar as possible, do not collect data from clients on their own caseloads.
- Conduct follow-up interviews on a small subset of interview subjects to check the accuracy of staff-collected data.

[29] This discussion is taken from Patton (2002), pp. 397-399.
[30] Expert consultants likely also would perform data analysis and report writing.

Questions to ask about data collection:

1. Does the choice of interview type — unstructured, semi-structured, or standardized — make sense given the type of information sought? For example, in an early-stage, exploratory study, one might expect to see less structured interviews than at a later stage when the specific information sought is clearer.

2. If a standardized interview protocol is to be used, does it contain enough open-ended questions to justify qualitative analytic techniques (and the smaller sample size), or would it make more sense simply to field a survey with a large enough sample to support quantitative analysis? Another way to think about this is to consider whether the open-ended items really explore important new ground or if enough already is known about the topic to formulate forced-choice questions.

3. Is the topic for focus groups likely to be controversial or otherwise sensitive enough that participants may not want to talk in front of each other? If so, a focus group approach may be inappropriate.

4. Do focus group participants share a similar relationship to the program? That is, are they all clients or line staff or supervisors? Clients may not be candid if program staff are present; line staff may not open up if supervisors are included in the same group. Generally speaking, focus group participants should be similarly situated with regard to the program under study.

5. Are focus group participants strangers to each other? Some experts recommend that focus group participants not have relationships outside of the focus group as prior relationships can alter the conversational dynamic.

6. What experience with conducting focus groups do the facilitators have? Strong group process skills are needed to encourage less verbal participants to speak and ensure that the conversation is not dominated by one or two individuals.

7. If observation is to be used, what experience/training do the observers have in conducting field work and preparing notes? How will data be captured (hand-written notes, audio recording, video recording), and does the budget adequately reflect time and costs for this activity?

8. What steps does the evaluator plan to take to address and minimize both types of researcher bias? Does the work plan reflect time for these activities?

9. If the proposal calls for data collection by program staff, what steps do the evaluators plan to take to ensure adequate training and oversight?

Qualitative Data Analysis

The goal of the analytic process in qualitative research is to tease out themes, patterns, and connections among ideas embedded in the data. Although some reduction in the volume of data produced by qualitative methods is necessary for a manageable analysis, it is important to bear in mind that preparing qualitative data for analysis should not simply quantify the data (Miles and Huberman, 1994). If evaluators collect qualitative data — such as responses to open-ended questions in an interview — but then during analysis simply extract categories of themes and report only the number of respondents who expressed that theme, the analysis is quantitative (even if the original data are qualitative). To be sure, the presentation of qualitative findings will include a certain amount of "counting" as a way to organize the findings and compare the frequency with which certain themes or patterns emerge in the data, but these numbers should not be the focus of the analysis.[31]

There are no formulas for analyzing qualitative data and fewer conventions than for quantitative analysis (Miles and Huberman, 1994; Patton, 2002). Reducing, organizing, indexing/coding, and displaying the data in various ways are the main activities undertaken in an effort to uncover key insights and patterns in the data. It is not unusual for analysis to begin while field work is still in progress; indeed, doing so may allow for necessary mid-course corrections and the development of preliminary findings.

Whatever analytic approaches are used, "analysts have an obligation to monitor and report their own analytical procedures and processes as fully and truthfully as possible" (Patton, 2002, p. 434). The process should be sufficiently transparent to allow for replication (Long, n.d.). Text Box 4 provides a brief description of the major components or steps in qualitative analysis that will give you a general idea about how the analytic process proceeds.

[31] As a purely practical matter, it's a waste of resources to collect rich, detailed interview data only to strip them down to a few numbers, discarding the rest. If the research questions can be answered well by this type of analysis, it may make more sense simply to field a well-designed survey and use quantitative measures. See also Hood (2006).

TEXT BOX 4 Overview of the Qualitative Data Analysis Procedures[32]

- **Preparing the data:** The goal at this point is to convert the messy, raw data into words presented in an intelligible format that can be read, edited, checked for accuracy, coded, and analyzed. This may take the form of verbatim transcription (interviews and focus groups), detailed narrative (field notes), or reduced and simplified summaries of these two. Summaries may also be the first step in analysis even if lengthier data collection, such as full transcripts, are produced. Studies with tight budgets and short time frames may choose to analyze the summary documents, rather than the longer ones.

- **Coding the data:** Coding is a form of indexing used to organize the massive amount of data produced by qualitative research. A critical step in data analysis, coding labels sections of text to describe or interpret key meanings and ideas expressed. Both verbatim or summary format data can be coded, and researchers often do more than one coding pass for each data document, possibly doing basic descriptive coding the first time, moving to more interpretive or pattern codes on a later pass.

  Codes are driven by the research questions. Many researchers will develop the list of codes to be used in the analysis prior to the beginning of field work.[33] Codes lists often change as field work and analysis proceeds; this is typical, not a sign that something is wrong. Today, researchers use special software, such as ATLAS.ti or NVivo to assign codes to sections of text that then can be more easily retrieved for additional analysis.

- **Preliminary efforts to synthesize and make sense of what is being learned:** As analysis proceeds, researchers undertake a variety of activities, either formally or informally, to begin to synthesize findings — and make course corrections as necessary — by producing memos, case summaries, or meeting with other research team members to discuss cases and impressions.

- **Producing data displays:** Although not all analyses take this step, Miles and Huberman (1994) argue that sorting and organizing segments of data into matrices or network diagrams allows researchers to compare ideas within or across cases and see connections among themes and ideas more readily than sequential reading of interview transcripts or observational narratives permits. Some version of these displays may also be useful for inclusion in the final report as a way to present and summarize key findings, as well as document the evaluators' process for arriving at conclusions.

Sources: Miles & Huberman, 1994; Patton, 2002

---

[32] This section is taken primarily from Miles and Huberman (1994), particularly chapters 4-9, and Patton (2002), chapter 8. For additional information, see Cohen and Crabtree (2006) at http://www.qualres.org/HomeComm-3821.html., and Boyatzis (1998).

[33] Grounded theorists, in contrast, generally wait until some of the field work has been completed, written up, and scrutinized before developing analytic codes. This approach may be seen less often in applied research or program evaluations, but you may encounter it.

Inter-Rater Reliability

A rigorous analytic procedure will include steps to ensure that the codes are being applied consistently. Inter-rater reliability refers to the level of agreement among two or more staff doing the coding.[34] Intra-rater reliability refers to how consistently any one staff person applies the codes over the course of the analytic process. The evaluation team may calculate a ratio indicating the level of agreement among and within coders.

Inter-rater reliability: The exact steps may vary slightly, but the general approach is to have all staff tasked with coding to start by coding individually the same portion of the data — perhaps the first few pages of a subset of interview transcripts, for example — and then comparing their work. Typically this process will reveal some dissimilarities in how specific blocks of text are coded. The team should respond by clarifying or expanding code definitions, perhaps adding or omitting certain codes, and revising the code list and definitions accordingly before beginning the actual work of coding the data (Boyatzis, 1998; Miles & Huberman, 1994; Patton, 2002). Additional staff training may also be indicated.

Intra-rater reliability: Over the course of a long analytic process, coding staff may inadvertently start using codes slightly differently by the end of coding. Miles and Huberman (1994) recommend that researchers check for intra-rater reliability to minimize this type of inconsistency. This procedure calls for having each staff person go back, perhaps when they are about two-thirds of the way through the coding process, and re-examine data coded early in the analytic process and ensure that they are applying codes to the text in a consistent manner.

Measures of consistency: Once the data are coded, a subset of coded documents may again be examined and measures of consistency among coders calculated (Boyatzis, 1998; Patton, 2002). Most typically, this measure is expressed as a percentage of agreement; that is, out of all the instances of coding (for example, blocks of text that got coded), how often did the coders agree? [35] There is not a single, agreed-upon threshold of acceptable consistency, and experts disagree: Miles and Huberman (1994) suggest that both inter- and intra-coder agreement should eventually be in the 90% range; Boyatzis (1998) sets the bar lower, at 70% or above. Projects with more complex coding schemes are likely to show lower levels of consistency among raters. What matters is that evaluators make an effort to assess and report levels of agreement among raters.

Questions to ask about qualitative data analysis:

Not all program evaluations will include all these procedures; time or budget constraints may limit analytic steps. But all studies need to process and code the data; how much time is spent on higher-order interpretation will vary. And some effort should be made to assess and demonstrate that analytic codes were applied consistently. There is not one right way to analyze qualitative data. What matters is that the evaluators be explicit about what they did. Given the variation and complexity of qualitative data analysis, the federal project officer (FPO)'s role may consist primarily of general oversight and requests for accountability and transparency.

---

[34] In many projects, different staff will ultimately code different portions of the data. However, in large studies with complex data coding schemes, Patton (2002) suggests that having every item of data coded by at least two staff, although labor-intensive, is the most rigorous approach.

[35] Boyatzis (1998) also suggests that for some types of codes, correlation coefficients may be used to assess inter-coder consistency. The reasons are technical and beyond the scope of this guidance, but you should be aware that some evaluators may prefer this method for assessing and reporting inter-rater reliability.

1. Has the evaluator explained exactly how data analysis will proceed (i.e., what data will be coded, who will do the coding, how many coding "passes," etc.)? There is not one right way to analyze qualitative data, but the evaluator should be able to describe a systematic approach that aims for consistent, unbiased findings, and that is thorough enough to allow another researcher to replicate the process.

2. If you decide to examine the code lists: Do the codes capture the constructs of interest? You may wish to ask to see the final coding scheme, at least for the major codes. Some studies may also prepare secondary and tertiary codes as well; how in depth you want to get may depend on your own expertise and time.

3. Because qualitative studies often take early data analysis steps while field work is still in progress, evaluators may produce interim reports of findings. If these are shared, does the direction of the analysis appear to be getting at the research questions? What questions are not being adequately answered? What plans does the evaluator have to correct this deficiency?

4. If interview data are to be summarized on a data capture form (rather than being fully transcribed), is this approach consistent with the goals of the study? That is, how likely is it that important aspects of participants' perceptions may be lost or the stories they tell less useful if responses are only summarized?

5. If verbatim transcripts are to be prepared, do project timelines adequately allow for this activity? Bear in mind, too, that coding verbatim transcripts (and in-depth field notes) likely will be more time-intensive than analysis of summary data.[36]

6. What procedures for ensuring inter-rater reliability were followed? How good was the final level of consistency among staff performing the coding?

Summary

Rigorous implementation of a sound research design, careful sampling, well thought-out data collection and capture, and methodical data analysis are essential if the program evaluation is to generate valid, useful findings. It should be clear from this overview that qualitative research, when properly carried out, is just as rigorous as is quantitative research. It should also be clear that implementing these methods well takes an evaluator who is experienced in their use: Good qualitative research takes training and experience, "just going out and talking with people" will not yield credible findings, and valid interpretations of the data must be based on rigorous analytic procedures. The final section of this document dives a bit deeper into research methods, particularly analysis and interpretation, to consider assessment of the credibility of research findings.

[36] As an example, a verbatim transcript for a one-hour interview will run to 10-15 single-spaced pages of text (Patton, 2002). Interview summary sheets typically are shorter and contain less information. They also may pre-organize responses by question which can speed analysis. See also the cost section of this document.

PART FOUR: THE CREDIBILITY OF QUALITATIVE FINDINGS

Although the discussion of credibility comes last in this document, it is a central goal that should guide the evaluation process from start to finish: The credibility of research data, and the conclusions derived from their analysis, begins with the implementation of sound, consistent methods for collecting and analyzing those data (Miles & Huberman, 1994; Patton, 2002; Silverman, 1990; Sofaer, 2002). The elements of rigorous qualitative methods, as well as the importance of engaging an experienced evaluator to ensure their implementation, have been addressed in previous sections of this document. Now let us turn to two additional building blocks upon which credible conclusions rest: Whether or not the data support the conclusions, and how widely the findings can be generalized.

First, a few words about terminology: Words such as validity, credibility, reliability, and trustworthiness all have been used to describe data and conclusions that accurately reflect the phenomena they purport to describe (Boyatsis, 1998; Miles & Huberman, 1994; Patton, 2002; Secrest & Sidani, 1995). Sometimes these terms are applied broadly; at other times, they refer to specific things such as how well measures reflect underlying constructs or whether the data support or rule out competing conclusions.[37]

To avoid confusion, this document will use these terms as follows: "trustworthiness" and "credibility" are the most general terms used to refer broadly to the goodness or accuracy of data or conclusions. The term "reliability" has been used previously in this document mainly to refer to what might more accurately be called "inter- and intra-rater reliability;" that is, the degree of consistency or agreement among two or more researchers — or the same researcher over time — analyzing the same set of data (see the data analysis section of this document for details).

The use of the term "validity" will roughly follow the convention established by Campbell and Stanley (1963) in that I will distinguish between two main types of validity. *Internal validity* refers to how well the data support the conclusions and specifically to how well alternative conclusions can be ruled out. *External validity* refers to the degree to which conclusions may apply to settings other than the one studied. However, because qualitative research does not always follow the same criteria for assessing validity as does quantitative research — and sometimes prefers to use different terms altogether — the variations in both terminology and meaning will be discussed in subsequent sections.[38]

Internal Validity: Ruling Out Alternative Explanations

However complete and trustworthy they may be, descriptive data do not speak for themselves; they must be interpreted for meaning and conclusions drawn in order to be useful. A critical concern with any research (qualitative or quantitative) is "whether the conclusions being drawn from the data are credible, defensible, warranted, and able to withstand alternative explanations" (National Science Foundation, 1997). The term "validity" often is used to refer to how well conclusions approximate reality. Some researchers — and this guidance — use the

[37] When referring to measurement instruments, particularly standardized instruments, "reliability" refers specifically to how consistently a construct is measured, and "validity" refers to how well the instrument captures the underlying construct it purports to measure.

[38] Campbell and Stanley were concerned chiefly with causal conclusions — that is, whether variable A caused variable B. Because qualitative research generally is not focused on causal explanations, this guidance is following Shadish (1995a) and applying the idea of validity to all types of conclusions, not only causal ones.

term "internal validity" to capture this concept (Cook & Campbell, 1979; Silverman et al., 1990). Qualitative researchers may also use the terms "credibility" or "authenticity" to refer to the goodness and accuracy of conclusions (Miles & Huberman, 1994). Whatever term is used, the general idea is the same: Valid conclusions in program evaluation provide a reasonably accurate portrayal of what actually is going on with the program under study.[39]

But for those familiar with the more rule-based conventions used to assess quantitative methods, the validity of qualitative research may be viewed with skepticism (Miles & Huberman, 1994). There are fewer analytic conventions with qualitative research, and therefore the role of the researcher as the interpretive gate-keeper is critical. Patton notes that a significant "barrier to credible qualitative findings stems from the suspicion that the analyst has shaped findings according to predispositions and biases" (2002, p. 553).[40]

However, there are strategies that qualitative researchers can and should employ to ensure that their conclusions present as full and accurate a picture of what is being studied as possible. Conclusions must be supported by the data, and the interpretive process needs to be rigorous. Rigor in interpretation rests on several strategies including considering rival conclusions, looking for negative or disconfirming cases, using triangulation, and getting feedback from study participants (Patton, 2002).

Rival Conclusions

As qualitative field work and analysis proceeds, the researcher begins to look for patterns in the data and to formulate explanations for why the data look as they do. Because qualitative research generates volumes of sometimes unwieldy data, it can be tempting to seize prematurely upon a likely explanation. When that happens, the researcher may, even inadvertently, begin to focus only on the bits of data that support the favored conclusion, dismissing other data as outliers or exceptions (Miles & Huberman, 1994; Patton, 2002). But the integrity of the analysis rests on the researcher resisting this temptation and instead looking actively for rival conclusions, for other stories that can be told to account for the data (Patton, 2002).

Generating rival hypotheses or conclusions should be done as data collection is proceeding, and it continues through the data analysis process (Miles & Huberman, 1994; Patton, 2002). One strategy that can help with formulating and exploring rival hypotheses is to look for cases that do not "fit" the favored explanation (see next section on contrasting and disconfirming cases) (Becker, 1970). Another is to make hypotheses explicit and share them among the evaluation team, soliciting feedback and inviting opposing ideas (Miles & Huberman, 1994).

It is likely that this process will not yield absolute support (or non-support), but instead will point to a conclusion that fits "the preponderance of data" (Patton, 2002, p. 553). Not all social researchers generate formal hypotheses to be tested, but during the research process, a

---

[39] Cook and Campbell (1979) remind us that references to the validity of research findings should always be understood as if prefaced with the words "approximately" or "tentatively" because one can never be completely certain as to what is true.

[40] There are some qualitative frameworks, such as social construction, that value subjectivity and view it as an avenue to achieving insight. That framework and its criteria for assessing credibility differ somewhat from what this document has presented because it is less typically found in applied government and policy research (Patton, 2002). However, it should be noted that qualitative researchers working within those philosophical frameworks are clear that the researcher's biases, assumptions, and preconceived ideas must be made explicit at the outset of research and as the project proceeds. The term "reflexivity" captures the idea of being conscious of one's own perspective and biases, and being reflexive is a critical component of constructivist interpretation.

researcher should at least informally be thinking about the different "stories" that the data could tell.

This process takes time, and the evaluation team needs to resist reaching a too-hasty conclusion, however tempting that may be. The implication for FPOs is that the work plans (and budget) allow ample time for interpretation and analysis (as has been stated elsewhere), and that the FPO asks questions about emerging conclusions and what evaluators are doing to seek alternative explanations. The evaluator should be able to report on the alternative hypotheses or explanations that were formulated, tested, and discarded — and why (Patton, 2002).

Contrasting and Disconfirming/Negative Cases

Once a set of patterns or hypotheses have emerged from the main body of data, the researcher should search for and analyze cases that *do not* fit the pattern or support the hypotheses (Becker, 1970; Miles & Huberman, 1994; Patton, 2002). The search for such cases may include those that refine or expand a hypothesis as well as those that directly refute initial tentative conclusions.

Contrasting cases may include those that are outliers or extreme cases. As patterns emerge, there may be some that apply to many cases. As with rival explanations, the researcher should not latch too hastily onto these mainstream patterns, ignoring those that do not conform. Instead, a concerted effort to locate cases that are outliers from mainstream experience or represent extreme views or experiences should be sought and explored. [41] Such cases force the researcher to revisit tentative conclusions and revise them in order to account not only for the mainstream pattern, but also for the non-conforming cases.

The search for *disconfirming cases*, sometimes referred to as negative cases or negative evidence, actively seeks disconfirmation, not simply refinement, of initial hypotheses and conclusions (Miles & Huberman, 1994). In this approach, once a researcher has reached a tentative conclusion, he or she asks whether any cases or informants oppose or contradict the conclusion and then aggressively looks through the data to find such cases and report on them. In some cases, exceptions may "prove the rule" as when, for example, negative participant experiences are discovered to be associated with poor implementation or with staff who had not been well trained. In such cases, the more specifically qualified conclusion that positive experiences can occur *if the program is implemented correctly by properly trained staff* would be strengthened (Miles & Huberman, 1994).

The write-up of the study should include how contrasting or disconfirming cases were explored and how they influenced the conclusions drawn. Not only will doing so shed light on potentially important processes, events, or subgroups, but it also will demonstrate the research team's integrity and credibility (Patton, 2002). The use of data displays can help minimize bias by keeping the range of findings in front of the researcher and making sure they all get considered in the analysis; some version of these displays can also be a valuable addition to the final report that will add to its credibility (Hood, 2006; Miles & Huberman, 1994).

Although much of the exploration of contrasting and disconfirming cases will happen outside your direct observation as FPO, you can ask questions about how the evaluator sought out and explored negative or disconfirming cases, and how the study's conclusions were (or were not) modified. You also can request that the final report include data displays reflecting the

[41] Obviously, this assume a well-thought through sampling design has been implemented from the start so that a diversity of cases will be included in the study (Miles & Huberman, 1994).

range of data informing key conclusions and that the process for accepting or rejecting these be described.

Triangulation

If a study has been able to implement a triangulation approach — that is, if it has collected data from different sources, using different investigators, or using different methods — conclusions can be strengthened by comparing the range of information obtained from independent sources and exploring any inconsistencies found (Miles & Huberman, 1994; Patton, 2002). There are several types of triangulation (see Text Box 5).

TEXT BOX 5: Triangulation

- Methods triangulation — Comparing data collected using different methods (i.e., observation, interviews, and document review). Some definitions include here using qualitative and quantitative data.[42]
- Data source triangulation — Comparing data collected from different sources (informants, places, times).
- Researcher triangulation — Comparing data collected by different investigators or conclusions reached by different analysts looking at the same body of data.
- Theory triangulation — Using multiple theories or perspectives to interpret data.[43]

Sources: Miles & Huberman, 1994; Patton, 2002.

A variant of triangulation is replicating findings across cases. In a multi-site program evaluation, this might consist of testing how well a conclusion generated from the data at one site "fits" with that found at another site (Miles & Huberman, 1994).

The goal of triangulation is not to hope all data yield exactly the same conclusion - that is not likely — but to *explore and understand inconsistencies* to gain better insight into what has been studied and so to arrive at more credible conclusions (Patton, 2002). The use of data displays by the researchers to show the range of data across data sources or methods can help illuminate comparisons and contrasts (Miles & Huberman, 1994). Finally, it should be clear by now that in order to use triangulation to strengthen the validity of conclusions, the study must have built in multiple methods, data sources, etc. from the beginning. Miles and Huberman regard triangulation as "a way of life," not simply a tactic (1994, p. 267). As an FPO, you can ask questions of evaluators, as the evaluation plan is being formulated, with an eye to obtaining data in multiple ways so that this validation strategy can be implemented. You can also look at

[42] Miles and Huberman (1994) prefer to distinguish the use and comparison of findings generated by using qualitative and quantitative data as "triangulation by data type" (p. 267). Although Patton considers comparing qualitative and quantitative data as a methods triangulation approach, the term is less important than the idea that data generated by different and independent data collection techniques can be explored for consistency.

[43] Patton includes getting feedback from informants under this type of triangulation; i.e., "triangulation of perspectives" (560-561). The present document will follow Miles and Huberman's classification and discuss informant or participant feedback as a matter separate from triangulation (see next section).

whether or not the work plan and budget will support triangulation and ask evaluators what trade-offs are being made if time or resources are limited.

Getting Feedback from Participants/Informants

One final strategy to confirming the validity of conclusions is to obtain feedback from informants and program participants. Such feedback may be sought at various points in the study; early products like interview transcripts may be shared with the interviewee for confirmation.[44] More often, later products would be shared, and here also there are options for doing so. Feedback may be sought on a short summary of findings and conclusions, such as an executive summary, or on the full report (Miles & Huberman, 1994; Patton, 2002). Data displays or other ways of providing a quick overview of the information are useful for showing how conclusions were reached, although the evaluators must take care with regard to sharing details on specific incidents that may be sensitive (Miles & Huberman, 1994). Finally, if an abstract or executive summary is specially prepared for sharing, the evaluator should make sure the language used is not overly technical, but is something that participants will understand.

The objective is not to have informants/participants agree with everything the report says or even with one another. That is not likely in any case. The hope is that informants are able to "relate to and confirm the description and analysis" (Miles & Huberman, 1994; Patton, 2002, p. 560). Seeking feedback should be regarded as a way to learn more about the program, provide context, or generate alternative interpretations (Miles & Huberman, 1994; Patton, 2002). If informants disagree with the findings, they may be prompted to offer additional information that might not otherwise have been forthcoming in order to clarify their objections, and that can help the evaluator better understand the program or refine the study's conclusions (Patton, 2002).

The evaluator's own integrity and judgment are critical here, as negative findings quite likely will be resisted by program staff and leadership. If the evaluators find they have overlooked critical information, that could prompt a revisiting of the conclusions, but they should must be sufficiently independent to stand by otherwise well-supported findings.[45] And the researchers should make sure informants understand in advance that they will not have final say over the content or conclusions of the report.

As with other validation techniques, seeking feedback from others takes time and effort. If the evaluators plan to use this approach, you can ensure that the work plan and budget allow adequately for these activities. At a minimum, it may mean having the evaluator generate an interim report in sufficient time for gathering feedback and then preparing a revised final report before the funding runs out.

Questions to ask about internal validity

Before listing some questions you can ask your evaluators to determine the validity of their conclusions, a couple of reminders are in order: First, as the foregoing discussion has suggested, so much of what needs to be done to strengthen validity will happen outside your direct observation. Because of this, the experience and credibility of the researcher and the

[44] Miles and Huberman (1994) caution that seeking feedback while data collection is ongoing, although potentially useful, does run the risk of introducing bias if informants react by altering their behavior or perspectives.

[45] There may be situations where getting feedback from participants and informants on an interim report is not advisable because of the potential to generate controversy that could bias the final report. The evaluator and FPO need to weigh the benefits and risks for each project.

evaluation team are critical and need to be considered when selecting an evaluator. Second, many of the techniques described in this section rely on rigorous methods having been implemented from the start in order to ensure a range and quality of data that will allow the comparison of cases, generation of rival hypotheses, etc. Once those two elements are in place, here are some specific questions you can discuss with the evaluator:

1. What procedures or rules did the evaluator use for the confirmation of hypotheses or conclusions? Are there remaining areas of uncertainty, and if so, what are they?

2. Can the evaluator discuss what rival conclusions were considered? What were they, and do they seem plausible to you? What led the evaluator to reject them?

3. What contrasting or negative evidence was explored and what was learned from these instances? Does the report use data displays to document the range of findings across cases or informants?

4. What efforts to incorporate triangulation were made by the evaluator? How did the findings compare across data sources, methods, and investigators? Where there is divergence, how does the evaluator account for that?

5. How has the evaluator sought participants'/informants' feedback on the initial report? What did the participants/informants say? How is their feedback reported and reflected in the final report?

External Validity: Generalizing the Findings

Once a study's conclusions have been accepted as credible, a logical next step is to ask if they can be generalized beyond the settings or individuals studied. The term "external validity" often is used to capture this idea; qualitative researchers may speak also of "transferability" or "fittingness" (Cook & Campbell, 1979; Miles & Huberman, 1994). The concept of generalization is subtle and complex — and much debated among both quantitative and qualitative researchers. A full discussion is well beyond the scope of this document. The objective of the present discussion is to help you think about what can be generalized from a given program evaluation and how well the study's methods support claims of generalizability. Let us start by clarifying a couple of very basic issues:

- Generalization does not consist solely of the sample-to-population claims most familiar to many researchers (and consumers of research). Rather, generalization goes to a broader question: What can be extrapolated from the study of a specific setting to other settings (Firestone, 1993; Patton, 2002)?
- Generalizations are not established "facts," but instead are claims or arguments that must be assessed critically and tested further (Firestone, 1993; Miles & Huberman, 1994). Patton calls extrapolations "modest speculations on the likely applicability of findings to other situations under similar, but not identical, conditions" (2002, p. 584).

Generalizations are tentative at best, and this holds true for both qualitative and quantitative research findings. For qualitative research in particular, claims of generalizability have never been strong (Firestone, 1993).[46] Some qualitative researchers are skeptical, or even dismissive, of generalization, preferring to focus on understanding specific phenomena as they interact with a specific context, rather than looking for typical or average outcomes across settings (Firestone, 1993; Miles & Huberman, 1994; Stake, 1978 as cited in Patton, 2002).

However, many other qualitative researchers recognize that the question of whether or not findings from a specific study have any applicability to other settings remains an important one that "does not go away" (Miles & Huberman, 1994, p. 173). When it comes to program evaluations, those who fund and use these studies quite reasonably want to know if the work yielded any knowledge or lessons that can be useful in future programs (Patton, 2002).[47] When the study includes findings based on qualitative data, the questions then are what, exactly, can be generalized from those findings and what determines the confidence we can have in those claims.

At the risk of over-simplifying a very complex topic, this remainder of this discussion will be organized using Firestone's (1993) three types of generalization in an effort to clarify and distinguish among the types of extrapolations researchers can make and to identify how well each type is supported by qualitative research: Sample-to-population extrapolation, case-to-case transfer, and analytic generalization. Each of these approaches establishes a basis or "argument" for making the leap from the setting studied to other settings or to theory, but those arguments come with qualifications and limits.

Sample-to-population extrapolation

This may be the most familiar type of generalizability, and it often is thought to offer the strongest support for generalizing claims (Firestone, 1993). Its strength rests on the use of large, random samples and probability theory. If a study looks at a random sample of urban 14-year olds, assuming rigorous research methods were implemented, the researchers could say, with a specific, quantified level of confidence that the findings should apply to the larger population of urban 14-year-olds. This type of probability-based, sample-to-population generalizability generally cannot be inferred from a study using qualitative methods, primarily because the type of sampling used with such studies — smaller, nonrandom samples — does not support probability-based population inferences (Firestone, 1993; Hood, 2006; Miles & Huberman, 1994).

[46] Some researchers question if generalizability is even an appropriate goal for qualitative research (Denzin, 1983; Guba & Lincoln, 1981, both as cited by Miles & Huberman, 1994). One argument put forth by some researchers is that because social phenomena (like social programs) are inherently local and context-specific, good qualitative work should focus first on getting the particulars right within a specific context (Guba & Lincoln, 1981; Stake, 1995, both as cited in Patton, 2002).

[47] Patton notes that summative evaluations — that is, those focused on participants' outcomes — most often rely on quantitative data. Qualitative data in these studies generally are used to "add depth, detail, and nuance to quantitative findings" (2002, p. 220). He goes on to observe that formative evaluations, which focus on shaping and improving a program or policy, and therefore are less concerned with generalizing beyond the setting under study, "often rely heavily, even primarily, on qualitative methods" (Patton, 2002, p. 220).

Case-to-case transfer

This extrapolation approach has origins in program evaluation: Practitioners in a given setting may wish to try out an approach that seems promising based on evaluation findings from another setting. Qualitative methods have long been associated with this type of generalization (Firestone, 1993).

Transferability from cases does not rest on large sample findings or probabilities, but on "rich, detailed, thick description" of the studied case that permits the reader to assess the match and determine how applicable the findings may be to the new setting. Firestone also notes that the inference of applicability is with *the reader*, not the researcher; the argument is that the researcher cannot know enough about non-studied cases to generalize. For this reason, some analysts regard case-to-case transfer as a comparatively weak basis for generalization (see for example, Kennedy (1979, cited in Firestone, 1993). However, Guba and Lincoln (1985, cited in Firestone, 1993) hold that reader inferences are "the only defensible strategy" (Firestone, 1993, p. 18). Regardless of whether generalization rests with the reader or the researcher, making such claims based on case-to-case transfer requires that investigators provide descriptive details about the setting that are sufficient to allow a determination of applicability to another setting (Firestone, 1993).

Analytic generalization

The third type of generalization Firestone identifies works to generalize findings to underlying theory. This type of extrapolation offers support for (but does not prove) the theory. Instead, these generalizations are treated as working hypotheses that need to be tested in new settings and contexts in order to determine their "analytic generality" (Firestone, 1993; Guba, 1978, cited in Patton, 2002; Miles & Huberman, 1994, p. 31). When a finding is generalized to theory, there certainly is the implication that the finding will hold across other settings than the one studied, but unlike sample-to-population generalization — where the extrapolation is from the sample to the population *from which the sample was drawn* — in analytic generalization, the theory may apply to other populations as well (Firestone, 1993). However, this basis for generalizing places the emphasis on knowledge and ideas, not cases and settings. Firestone regards this type of generalizability as particularly promising for qualitative research (1993).

The strength of claims generated through analytic generalization rests on sampling for diversity across theoretically important constructs or exploring multiple sites. The researcher also needs to tease out and specify what the important factors or constructs are (Firestone, 1993). Where a pattern has emerged in the data — and particularly if it emerges across a variety of cases (i.e., individuals, settings, etc.) — the researchers can have more confidence that the pattern is theoretically important and may apply to situations other than the one that was studied (Miles & Huberman, 1994). Performing a comparison and contrasting of cases, what Miles and Huberman refer to as cross-case analysis, is an important analytic technique researchers can use to increase confidence that findings will have applicability in settings other than the one studied. In particular, when findings hold true across contrasting cases, not only is internal validity strengthened (as described in the previous section), but confidence that the findings may make sense in other settings also is increased (1994).

Maybe an example will help clarify what is, admittedly, a subtle and somewhat academic point: Miles and Huberman use the example of a study of how role modeling socializes young children in which only a few kindergarten classrooms could be studied. Assuming the sample was drawn to maximize the researchers' ability to look at a variety of circumstances and events

— teachers of different gender, teachers' with a range of inclinations toward socialization versus academics, observation of a range of classroom events (story-time, show-and-tell, recess, etc.) — the findings likely would yield some insights that apply to theories about how role modeling works (1994, p. 27). These may also apply to all kindergartens in the population sampled — or even to other, similar classrooms — but the argument is not advancing this claim explicitly.[48] The argument, instead, focuses on the generalization of knowledge to theory, not specifically to what populations it might also apply.

Summary

The distinctions among types of generalizability may be difficult to grasp, so let us try (again, at the risk of over-simplification) to summarize the key points:

- Generalizability is a broad concept that applies to the question of what knowledge or understandings can be extrapolated from what was studied to other, not-studied settings.
- There is more than one type of generalization, and each one makes a slightly different argument:
    - *Sample-to-population extrapolation*: Argues the findings will apply to the population from which the sample was drawn. Rests on large samples and probability theory.
    - *Case-to-case transferability*: Argues that the findings will apply to sufficiently similar cases. Rests on detailed, thick description.
    - *Analytic generalization*: Argues that the findings apply to (and perhaps expand or refine) existing theory (which may in turn apply to populations not studied). Rests on diverse samples and specification of which factors influence findings.
- All claims of generalizability, including those generated by large, random sample studies, are tentative and need to be verified by further study.

Finally, to quote Firestone: "[Q]ualitative methods should not be avoided because of the fear that their claims for broad relevance are especially weak. That is not the case" (1993, p. 22).

Questions to ask about generalizability

Once again, it is critical that rigorous sampling and analytic techniques have been implemented right from the start of the evaluation in order to support generalizability. Here are a few questions to ask of and discuss with your evaluator; most of them are taken or adapted from Miles and Huberman (1994, p. 279).

1. How do the sampling plan and analytic approach incorporate strategies that will strengthen the generalizability of the study's findings?

2. Does the final report discuss what reasonably can be generalized from this study? On what basis is the evaluator claiming generalizability?

[48] You could argue, and Firestone (1993) might well agree, that the claim is implicit.

3. Does the report explore how the sampling or analytic approach may have limited generalizability and in what ways?

4. Which processes and outcomes described in the report are generic enough that they may be applicable in other programs and settings?

5. Have similar findings been replicated in other studies? Are the findings in agreement with or otherwise confirmatory of prior theory?

6. Does the final report suggest other settings that might provide a useful test of the findings?

CONCLUSION

> *"Applying guidelines requires judgment and creativity. Because each qualitative study is unique, the analytical approach used will be unique. Because qualitative inquiry depends, at every stage, on the skills, training, insights, and capabilities of the inquirer, qualitative analysis ultimately depends on the analytical intellect and style of the analyst. The human factor is the greatest strength and the fundamental weakness of qualitative inquiry and analysis — a scientific two-edged sword."* (Michael Quinn Patton, 2002, p. 433)

If you have finished reading this document, the above quote may not surprise you. Qualitative research can be quite rigorous, but it is not as structured as is quantitative research. That makes your job overseeing evaluations that include qualitative methods a bit harder, as there are few hard and fast rules to follow. What this document is intended to do is equip you with a better sense of how qualitative research is conducted and suggest questions to discuss with your evaluators so you can ensure the evaluation meets your agency's goals.

REFERENCES

Administration for Children and Families. 2012. *Evaluation Policy*. Available online at: http://www.acf.hhs.gov/programs/opre/resource/acf-evaluation-policy

American Evaluation Association. 2013. *An Evaluation Roadmap for a More Effective Government.* Washington, DC: Author. Online at: http://www.eval.org/p/cm/ld/fid=154

Becker, Howard S. 1970. *Sociological Work: Method and Substance*. New Brunswick, NJ: Transaction Books.

Becker, Howard S. 1998. *Tricks of the Trade: How to Think About Your Research While You're Still Doing it*. Chicago: University of Chicago Press.

Boyatzis, Richard E. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development*. Thousand Oaks, CA: Sage Publications.

Campbell, Donald T. & Julian C. Stanley. 1963. Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook for Research on Teaching*. Chicago: Rand McNally. (Also published as *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966. Online at: http://moodle.technion.ac.il/pluginfile.php/367640/mod_resource/content/1/Donald_T._%28Donald_T._Campbell%29_Campbell,_Julian_Stanley-Experimental_and_Quasi-Experimental_Designs_for_Research-Wadsworth_Publishing%281963%29%20%281%29.pdf)

Centers for Disease Control and Prevention. 2013. *Developing an effective evaluation report: Setting the course for effective program evaluation*. Atlanta, Georgia: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Division of Nutrition, Physical Activity and Obesity. Online at: http://www.cdc.gov/eval/materials/Developing-An-Effective-Evaluation-Report_TAG508.pdf

Cohen, Deborah. & Benjamin Crabtree. 2006. *Qualitative Research Guidelines Project*. Online: www.qualres.org/index.html.

Clark, Helene & Andrea A. Anderson. 2004. *Theories of Change and Logic Models: Telling Them Apart*. Presentation at the American Evaluation Association, Atlanta, GA. November 20014. Online at: http://www.theoryofchange.org/wp-content/uploads/toco_library/pdf/TOCs_and_Logic_Models_forAEA.pdf.

Cook, Thomas D. & Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.

Creswell, John W., Ann Carroll Klassen, Vicki L. Plano Clark & Katherine Clegg Smith. 2011. *Best Practices for Mixed Methods Research in the Health Sciences*. Bethesda, MD: National Institutes of Health, Office of Behavioral and Social Sciences Research. Online at: http://obssr.od.nih.gov/mixed_methods_research/

Festen, Marcia & Marianne Philbin. 2007. *Level Best: How Small and Grassroots Nonprofits Can Tackle Evaluation and Talk Results*. San Francisco: Jossey-Bass.

Firestone, William A. 1993. Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher*, 22(4) 16-23. Online at: http://mkoehler.educ.msu.edu/hybridphd/hybridphd_summer_2010/wp-content/uploads/2010/06/Firestone-1993.pdf

Fixsen, Dean L., Sandra F. Naoom, Karen A. Blase, Robert M. Friedman & Frances Wallace. 2005. *Implementation Research: A Synthesis of the Literature*. Tampa, FL: University of South Florida.

Framework Workgroup. 2014. *A Framework To Design, Test, Spread, and Sustain Effective Practice in Child Welfare*. Washington, DC: Children's Bureau, Administration for Children and Families, U.S. Department of Health and Human Services.

Guba, Egon G. 1978. *Toward a Methodology of Naturalistic Inquiry in Educational Evaluation*. CSE Monograph Series in Evaluation No. 8. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles. Online: http://www.cse.ucla.edu/products/monographs/cse_monograph08.pdf

Hood, Jane C. 2006. Teaching against the text: The case of qualitative methods. *Teaching Sociology*, 34: 207-223.

James Bell Associates. 2009. *Evaluation Brief: Common Evaluation Myths and Misconceptions*. Arlington, VA: Author.

Lincoln, Yvonne S. & Egon G. Guba. 1985. *Naturalistic Inquiry*. Newbury Park, CA: Sage Publications.

Long, Andrew. n.d. *Evaluation Tool for Qualitative Studies*. School of Healthcare, University of Leeds. Online at: http://usir.salford.ac.uk/12970/1/Evaluation_Tool_for_Qualitative_Studies.pdf

Miles, Matthew B. & A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook. 2nd Edition*. Thousand Oaks, CA: Sage Publications.

Morse, Janice M. 2003. A review committee's guide for evaluating qualitative proposals. *Qualitative Health Research*, 13(6): 833-851. Online at: http://www.sagepub.com/bjohnsonstudy/articles/Morse.pdf

National Research Council and Institute of Medicine. 2002. *Community Programs to Promote Youth Development*. Eccles, Jacquelynne & Jennifer Appleton Gootman (eds.). Committee on Community-Level Programs for Youth, Division of Behavioral and Social Sciences Education. Washington, DC: National Academies Press.

National Science Foundation. 1997. *User-Friendly Handbook for Mixed Methods Evaluations*. NSF Publication No. 97153. Frechling, Joy & Laure Sharp (Eds.) Westat. Online at: http://www.nsf.gov/pubs/1997/nsf97153/start.htm

Office of Behavioral and Social Sciences Research, National Institutes of Health (1999). *Qualitative Methods in Health Research: Opportunities and Considerations in Application and Review*. Bethesda, MD: Author. Online at: http://obssr.od.nih.gov/pdf/qualitative.pdf

Office of Planning, Research and Evaluation, Administration for Children and Families. 2010. *The Program Manager's Guide to Evaluation, Second Edition*. Washington, DC: Author. Online at: http://www.acf.hhs.gov/sites/default/files/opre/program_managers_guide_to_eval2010.pdf

Patton, Michael Quinn. 2002. *Qualitative Research and Evaluation Methods. 3rd Edition*. Thousand Oaks, CA: Sage Publications.

Puddy, Richard W. & Natalie Wilkins. 2011. *Understanding the Evidence, Part 1: Best Available Research Evidence. A Guide to the Continuum of Evidence of Effectiveness*. Atlanta, GA: Centers for Disease Control and Prevention.

Rist, Ray C. 2000. Influencing the policy process with qualitative research. In Denzin, Norman K. & Yvonne S. Lincoln (eds.), *Handbook of Qualitative Research*, pp. 1001-1017. Thousand Oaks, CA: Sage Publications.

Schorr, Lisbeth B. & Frank Farrow. 2011. *Expanding the Evidence Universe: Doing Better by Knowing More*. Washington, DC: Center for the Study of Social Policy.

Secrest, Lee & Souraya Sidani. 1995. Quantitative and qualitative methods: Is there an alternative? *Evaluation and Program Planning*, 18(1): 77-87.

Shadish, William R. 1995a. The logic of generalization: Five principles common to experiments and ethnographies. *American Journal of Community Psychology*, 23(3): 419-428.

Shadish, William R. 1995b. Philosophy of science and the quantitative-qualitative debates: Thirteen common errors. *Evaluation and Program Planning*, 18(1): 63-75.

Silverman, Myrna, Edmund M. Ricci & Margaret J. Gunter. 1990. Strategies for increasing the rigor of qualitative methods in evaluation of health care programs. *Evaluation Review*, 14(1): 57-74.

Sofaer, Shoshanna. 2002. Qualitative research methods. *International Journal for Quality in Health Care*, 14(4): 329-336. Online at: http://intqhc.oxfordjournals.org/content/14/4/329.long.

Tribal Evaluation Workgroup. 2013. *A Roadmap for Collaborative and Effective Evaluation in Tribal Communities*. Children's Bureau, Administration for Children and Families, U.S. Department of Health and Human Services.