CHAPTER TWENTY-TWO

QUALITATIVE DATA ANALYSIS

Delwyn Goodrick, Patricia J. Rogers

Most evaluations include some qualitative (non-numeric) material, such as open-ended responses in questionnaires, interview data—in the form of transcripts, notes, audio or video tapes—and textual data from existing documentation, such as policy documents and meeting minutes. Many evaluations also include visual types of qualitative data, such as photographs and images, and newer forms of qualitative products include materials generated in social media, such as tweets and posts. Some evaluations use solely qualitative data to draw conclusions, some use qualitative data in a supplementary role to complement quantitative data, and some give equal weight to qualitative and quantitative data in an integrated mixed-method approach.

Qualitative data allow readers to gain an understanding that goes beyond descriptive and inferential statistics. For example, although it may be useful to incorporate a count of how many young people in a mentoring program gained new skills, a more comprehensive understanding of the program is achieved from paying attention to how the young people describe their experience of learning new skills, and the way they connect their participation in the program with these outcomes. Qualitative data can provide deep insights into how programs or policies work or fail to work and more compelling accounts of the reasons for success and failure.

The value of qualitative data in program evaluation is well established. It can be challenging, however, to know what to do with the data generated from qualitative approaches, particularly for the evaluator who has little experience with qualitative data analysis. Within the constraints of tight timelines, evaluators must make sense of the material, analyze and synthesize it, and communicate the findings to others in ways that are clear, informative, credible, and useful.

This chapter provides an overview of four different ways to analyze qualitative data and provides advice about how to choose methods of analysis that address the purpose of the evaluation. It provides particular advice on how to code qualitative data in ways that are suit the type of analysis needed.

Types of Evaluation and Analytic Purpose

Evaluations are organized around key evaluation questions, and a particular form of analysis may address one or more of these questions. Qualitative data can be used across the full spectrum of evaluation types, such as needs assessment, program theory, process, and impact evaluation. While less common, qualitative data can be used in performance monitoring and economic evaluation.

Qualitative data can complement quantitative data, or can stand alone. Table 22.1 outlines the respective use of qualitative and quantitative data according to the type of evaluation being conducted.

Coding Data

Qualitative data analysis often involves sorting data into categories and labeling the categories. These categories are sometimes referred to as codes, nodes, or tags.

The coding process can be done in various ways, from structured processes with codes assigned a priori and applied to the data, to emergent coding frameworks that identify codes directly from the data. There are many options that may be undertaken in coding. Evaluators must think critically about these options and make reasoned choices. Particular methods of analysis may require formalized coding strategies. In later sections of this chapter we talk about different ways of coding data that suit the type of analysis being undertaken

Box 22.1 provides a definition of the distinctions we make between coding and categorizing processes.

TABLE 22.1. USES OF QUALITATIVE AND QUANTITATIVE DATA IN DIFFERENT TYPES OF EVALUATION.

Type of Evaluation	Common Uses of Qualitative Data	Common Uses of Quantitative Data
Needs Assessment/Situation Analysis (including strengths as well as needs)	In-depth understanding of people's experiences and needs Evocative description of people's situations Identification and documentation of problems and resources through photography or stories	Measuring the extent of needs and resources through social indicators and population surveys
Program Theory	Stakeholders' mental models of how programs or policies work Qualitative evidence (discrete or synthesized) from impact evaluations about effective programs and causal mechanisms	Quantitative evidence (discrete or synthesized) from impact evaluations about effective programs and causal mechanisms
Process Evaluation	Documenting the processes and implementation context of an innovation so it can be reproduced Periodic discussion and reflection on how things are going and how they might be improved Checking compliance with best practice/ quality standards using expert or peer review	Checking compliance with best practice/ quality standards using standardized measures of standardized processes
Performance Monitoring	Collection and review of qualitative performance measures Monitoring qualitative indicators over time	Collection and review of predetermined quantitative performance measures and indicators
Impact Evaluation	Exploration of a range of impacts (including unintended and unanticipated) and what they mean to participants Research designs that use historical information for causal attribution Values clarification, concept mapping of what is valued Realist synthesis of diverse evidence	Standardized impact measures and experimental and quasi-experimental research designs Statistical meta-analysis of effect sizes

TABLE 22.1. USES OF QUALITATIVE AND QUANTITATIVE DATA IN DIFFERENT TYPES OF EVALUATION. ***(Continued)***

Type of Evaluation	Common Uses of Qualitative Data	Common Uses of Quantitative Data
Economic Evaluation	Documenting the costs and benefits incurred by all groups (including negative outcomes) that cannot be quantified or monetized. Exploring what is valued by different groups	Quantifying costs and benefits and calculating the ratio of costs to benefits, or the cost for achieving a given level of effectiveness.

Box 22.1. Coding vs. Categorizing

- A **code** is a descriptive word or phrase that is intended to describe a fragment of data. For example, a segment of interview text could include multiple codes, participant role, length of time in the program, excitement about the program.
- A **category** is formed from the review of a cluster of codes. A category incorporates a collection of codes that relate to the same issue, topic, or feature in the data. For example, coded segments that relate to a participant's view of the program could be clustered together and examined. They could be further categorized as positive perceptions and negative perceptions and information sorted into each category. Categories can be theory led and developed a priori to be applied to the data or generated from an examination of the data (emergent or empirical categories). In evaluation, there is likely to be a combination of a priori and empirical categories based on the program theory or conceptual framework of the program being evaluated
- A **theme** is the outcome of categorizing and reflection by the evaluator on salient patterns in the data. Themes can be identified a priori (theory led) or inductively generated from examination of the data corpus. Themes are not a restatement of the key questions asked in an interview. "Perceptions of the program" is not a theme, and it is misleading to indicate that thematic analysis was undertaken if the evaluator has merely organized data under headings. Themes are often at a higher level of abstraction referring beyond what participants' stated to the meanings derived from the comparison of participant interviews. For example, "being invisible" may be identified as a theme that was drawn from an analysis of participants' stories or patterns noted in the data, but it may be that no participant actually used the term "invisible."

Analysis, including the coding processes that form part of it, requires hard thinking and dedicated space and time. It is difficult to retain the focus you will need if your analytic efforts are constantly being interrupted. Set clear time aside for coding and analysis. You will be more likely to have analytic breakthroughs when you persevere with the hard cognitive work of coding and analysis and limit interruptions.

Analysis and writing go hand in hand, so begin writing early. It may be useful to develop a working draft of the key evaluation questions and the relevant material/data that will address these questions. It is important to go back to the original data collection matrix to re-orient yourself to the intended purpose of the interviews (or other qualitative material that have been collected) and their relationship to the key evaluation questions.

The questions below (adapted from Baptiste, 2001) may assist decisions about coding and categorization processes.

Coding

1. When will I begin coding? Will I collect all the data and then start coding, or will I code as I collect data?
2. Will I use a software package or code manually?
3. Would it be useful to develop a number of categories (based on key evaluation questions or program theory) to assist in initial classification before I begin analysis, or will I rely on developing categories from the data as I review it?
4. Will it be useful to code by proceeding through each transcript or record, or would it be preferable to code in parallel (classifying and comparing responses to the same question across transcripts or records, or across stakeholder groups?
5. How detailed should the coding be? Will I code every data element that seems of interest or only code those directly related to the evaluation question.
6. How will the reliability of team coding be assessed? Will inter rater reliability be calculated or, will we adopt more interpretive processes to discuss and review coding structures?

Categorizing

1. How will I cluster codes into categories?
2. Is there an appropriate balance of categories? Is the level of categorization useful in informing analysis rather than being overwhelming?
3. Have I classified all the relevant data, or been overly selective in what I have classified?
4. Is each category linked to data elements and codes? Is there evidence for classification as a category? What are my decision rules?
5. Do my memos link codes and categories back to the context?

6. How will I move from categories to themes? How will I ensure I adequately link categories and themes to ensure that the written account is not fragmented?

Overview of Qualitative Analytic Methods

Across the wide range of methods for analyzing qualitative data, we have identified a useful heuristic to classify different methods. Each method of analysis has a primary focus and purpose. *Enumerative methods* turn qualitative data into numbers by sorting the data into a coding framework, tallying the data, and then developing categories. *Descriptive methods* show how concepts and entities are related by displaying them in tables or diagrams. *Hermeneutic methods* focus on identifying and eliciting manifest and latent meanings in the data. *Explanatory methods* generate and test theories about cause and effect relationships. An evaluation might use a number of these methods for analyzing different data or for analyzing the same data in different ways.

As the use of qualitative approaches has grown, there has been a corresponding growth in the nomenclature used to define and describe the methods, and variants in application and use of the methods across disciplines. The boundaries between the methods is not sharp, but we believe that the classification is helpful in making sense of the variety of potential strategies to analyze data, and that this classification can inform decisions about the most appropriate ones to address evaluation questions.

Table 22.2 shows examples of methods in each group according to purpose; methods in bold are discussed in more detail in later sections of this chapter. While there is overlap of methods across purposes, for example matrix displays may also be used for hermeneutic and explanatory purposes, we classify the methods by their primary use as documented in the literature and in evaluation studies.

In the sections that follow, we discuss the purpose of each method, and provide a description of an example that uses a particular method. We provide more detail about hermeneutic methods and explanatory methods. We focus on these, as hermeneutic methods are a common priority in qualitative evaluation contexts. Explanatory methods are also highlighted, as qualitative comparative analysis is gaining popularity in program evaluation to address causal evaluation questions, but there is limited information on the method in evaluation contexts.

Enumerative Methods

Overview Enumerative methods focus on categorizing qualitative materials so that they can be analyzed quantitatively. As for any analysis method, these

TABLE 22.2. FOUR GROUPS OF METHODS FOR ANALYSIS OF QUALITATIVE DATA.

Primary Purpose	Description	Examples of Methods
Enumerative	Summarizing data in terms of discrete and often a priori categories that can be displayed and analyzed quantitatively	**Classical Content Analysis (Krippendorf, 2013)** Word count Cultural domain analysis (pile sorts, free lists) (Spradley, 1980) Ethnographic decision models (Gladwin, 1989)
Descriptive	Describing how concepts and issues are related	**Matrix Displays (Miles, Huberman, and Saldana, 2014)** Timelines Concept maps/mind maps (Trochim, 1989) Template/framework analysis (Crabtree and Miller, 1999; Ritchie and Spencer, 1994)
Hermeneutic	Identifying or eliciting meanings, patterns and themes	**Thematic Analysis (Boyatzis, 1998)** Constant Comparative method (Strauss and Corbin, 1998) Thematic narrative analysis (Riessman, 2008) Framework analysis (Ritchie and Spencer, 1994) Discourse analysis (Wetherell, Taylor, and Yates, 2001) Qualitative content analysis (Schreier, 2012)
Explanatory	Generating and testing causal explanations	**Qualitative Comparative Analysis (Ragin,** 1987**)** Process Tracing (Collier, 2011)

methods are only appropriate for particular purposes and for particular types of data. While these methods can highlight patterns in data, they are a good option for analysis of open-ended survey questions and/or analysis of existing documentation or program records. This method can be used to analyze existing data (such as project documentation) or data created for an evaluation (such as transcripts of interviews). Enumeration is not recommended when the data is rich or detailed, as it tends to be overly reductionist and can decontextualize or distort the meaning.

Example Enumerative Method: Classical Content Analysis Classic content analysis is an example of an enumerative method of analysis, often used to analyze existing textual material such as newspaper reports or social media. Content analysis can involve examining the data for:

- Presence of absence of particular words, phrases, concepts, indicative of knowledge or awareness,
- Frequency in which an idea or concept is used, which may indicate importance or emphasis,
- The number of favorable and unfavorable words assigned to the idea, which may be evidence of attitudes about the program,
- The nature of qualifications made in the text that may point to intensity or uncertainty of beliefs and motivations, and
- The frequency of co-occurrence of two ideas, which may indicate associations between particular concepts or ideas. (Krippendorf, 2013)

The key steps to classical content analysis are presented below.

1. Clarify the purpose for using content analysis.
2. Determine the population and sample of texts/documents that will be analyzed. Sometimes it is possible to access and analyze an entire population (for example, all documented policies, or all tweets with a particular hashtag) and other times a random or purposeful sample is used.
3. Identify the unit of analysis to be categorized. The unit can be an entire document or a section of a document, depending on how it is to be analyzed. For example, is the purpose to identify what percentage of policy documents referred to a specific issue? Or to count how many times a particular policy issue was referenced?
4. Develop an initial coding framework This may be deductively driven (categories are developed from an existing conceptual framework derived from previous research or policy documents) or inductively generated (categories are identified from text statements present in the material). An initial draft can be developed and then modified as new issues arise when coding and categorizing. Formal decision rules for coding should be developed to ensure common definitional criteria are adopted and to maximize the reliability of coding processes. It may be helpful to develop a codebook for this purpose.
5. Test and revise the coding system. In classical content analysis inter-rater reliability can be calculated for teams involved in coding. Each member of the team is asked to code segments of text, and the coded segments are then compared for consistency across team members or raters. Further

training or testing is performed until sufficient reliability is attained. While there is no established standard of reliability in content analysis Krippendorf suggests that tentative conclusions are possible with a reliability coefficient between .67 and .80. In more interpretive approaches to inter-coder reliability, a co-efficient need not be calculated, but agreements and disagreements in coding should be discussed and addressed. This process of checking the reliability of coding is known as consensual coding.

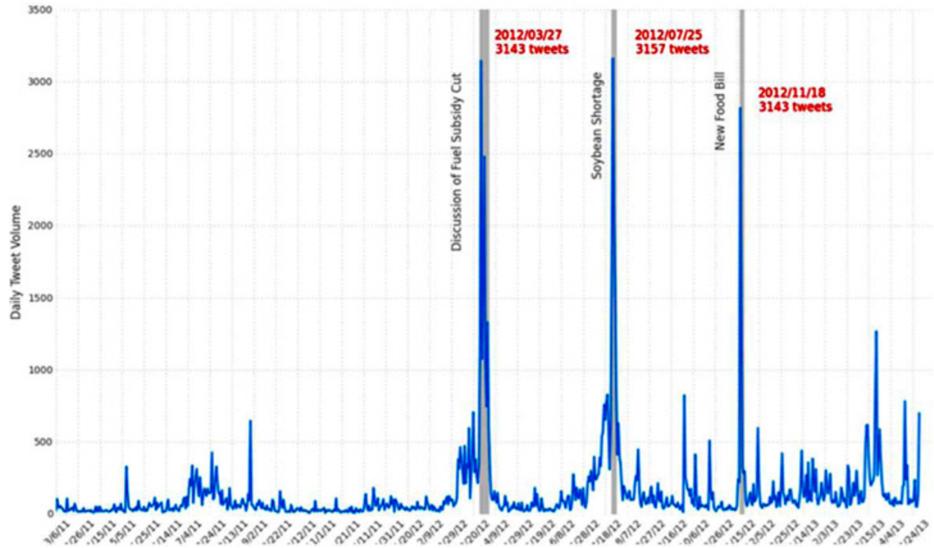
6. Code the data set. Apply the coding scheme to the data, identifying any issues arising where the codebook has to be modified. While the coding framework informs coding and classification, it is important to be open to other codes that emerge from the subsequent review of materials.
7. Summarize and present the findings. Once the data have been satisfactorily coded, the frequencies can be analyzed using various quantitative analysis methods, including frequency tables, graphs, univariate, bivariate, and multivariate statistics. If a random sample has been drawn, and an adequate response rate achieved, inferential statistics can be used to infer the characteristics of the population of documents, people or sites from which the sample has been drawn.

Good reporting of enumerative analysis includes transparency about the categories that have been used, and documentation of what has been coded under each category.

Application

An experimental project by the United Nations Global Pulse Lab, in conjunction with the Indonesian government, UNICEF, and the World Food Program, explored whether tweets could be used to provide real-time warnings of increases in food prices (UN Global Pulse, 2014). The dataset included all publicly available tweets from March 2011 to April 2013 (Figure 22.1), which were mined using the social media monitoring platform, ForSight. Tweets in Bahasa, Indonesia, were selected and then screened to only include those that included the keywords “harga” (price) and “naik” (rise) along with one or more food commodities. These 113,386 tweets were then classified as being either positive, negative, confused, or no emotion, initially manually and then by an algorithm. The monthly totals were compared to the official prices (for calibration) and the daily totals were graphed and compared to historical events and information from key informants. The trial showed that the tweets provided reliable evidence of increases in beef and chicken but not onions—giving agencies early warning of potential crises in food security.

FIGURE 22.1. DAILY TWEETS ABOUT FOOD PRICE RISES (MARCH 2011–APRIL 2013).



Enumerative analysis of the data was appropriate for this example because the purpose was to understand frequency, and the comparisons with official prices and local information showed that the data could usefully be analyzed in this way. A small set of keywords consistently identified the tweets of interest, and the high usage of Twitter among Internet users in Indonesia (Semiocast, 2012) meant there was adequate data about changes in food prices

Other trials have not always been so successful. Whereas initial analysis of reports of influenza symptoms on twitter had shown good correlation with actual reported cases, and provided early warning before official data were available (Paul and Dredze, 2011), later analysis of Google searches for the Google Flu Trends website overestimated incidence because the algorithms did not take account of heightened media coverage of flu outbreaks, so that searches were a poor indicator of incidence (Claburn, 2013).

When These Methods Are Appropriate

Enumerative methods are appropriate when the purpose of the analysis is to understand frequency. The sampling strategy and data collection process needs to have been designed for this purpose. These methods are also useful for categorizing open-ended responses on surveys. Frequency counts of the most or least identified topics will provide an indication of the prevalence of a topic. The evaluator still needs to pay attention to issues of survey quality, ensuring that the analysis is combined with a discussion of response rate and proportion of respondents that provided open ended responses.

Enumeration is not appropriate for analyzing data from a focus group discussion, where the purpose has been to explore different dimensions, not to calculate frequencies. Even reporting results in terms of "most people said." is not appropriate when the sample has been drawn for maximum diversity not for representativeness.

The categories must be able to be consistently and defensibly defined. Conclusions about what is the most common response will depend in part on how the responses are categorized. A broader category, for example, coding Greek and Italian responses under European, will appear to have a higher frequency than more specific categorization, for example by country. The evaluator must be confident about having coded the text accurately, which may require existing cultural knowledge or a review of the coding by content experts.

Enumerative approaches will be suitable if the evaluator has a large amount of qualitative data and clear categories for enumeration. The method can be persuasive to audiences interested in the salience and frequency of particular comments. The clear decision rules about categorization and the identification of discrete labels for enumeration mean the process is replicable.

Enumerative methods may also be potentially useful when the evaluation has generated thin data across a number of dimensions. In some cases, enumerative techniques may be the only possible option. For example, Del was asked to analyze interviews conducted by doctors with other doctors who had completed a quality improvement initiative. Following a review of the data, it was clear that most of the questions were closed-ended (for example, Did the initiative improve your practice?) or scaled (To what extent do you think that. . .?), to which the interviewee had responded with dichotomous answers. The only reasonable option for analysis was to count the frequencies of responses to particular questions and groups of questions and display these with illustrative quotes, where these were available. We believe that it would be unwise to rely on these methods if the evaluation data is rich and detailed, for example, when you have collected narratives or stories of change or have lengthy descriptive interviews, and your purpose is to elaborate the meanings.

The use of enumeration of qualitative data can be controversial in evaluation contexts. It is sometimes used when it is not appropriate—and some qualitative evaluators do not see them as being appropriate at all for analyzing qualitative data. It is understandable therefore that the conditions for their use need to be clearly assessed and justified for their use (see, for example, Patton's [2014] rumination 'Keep qualitative data qualitative').

One of the cautions when using these approaches is that they presume decisions about the frequency and concordance of words/terms can be determined by the evaluator. There may be little opportunity to check the meaning with evaluation participants. Counting words or terms can also be misleading, as participants will have a range of ways of expressing their views and use vocabulary that fits their intention, but that may be interpreted very differently by the evaluator. For example, if a participant describes a program as "wicked," does this communicate enthusiastic approval or strong disapproval, and does the participant's age suggest the meaning intended? It may require specialist knowledge to understand the intended meaning of participant statements. If this knowledge is not present in the evaluation team, some form of checking will be needed. (We discuss member checking or respondent validation in a later section.)

Evaluators using enumerative methods of data analysis tend to privilege quality criteria that are associated with traditional scientific criteria such as inter-rater reliability of coding of units that will be enumerated and independence of coding and assessment process.

Descriptive Methods

Overview. Descriptive methods of analysis focus on summarizing the information into a form that can then be compared and contrasted. We acknowledge that there is no such thing as pure description, but the function and form of these types of methods is to provide a strong descriptive account of materials gathered during an evaluation. While enumerative methods are descriptive, our classification of these methods as descriptive indicates that the intended focus and product is not enumeration.

Description is at the heart of good analysis and interpretation. As evaluators we must adequately describe the available data that has formed the basis of evaluative interpretations, conclusions, and judgments. Descriptive methods can be valuable for recording and analyzing fieldwork observations and interview and focus group data.

Example of Descriptive Method: Matrix Displays. A common analytic choice often associated with descriptive methods in evaluation is matrix displays.

Matrix displays can be used for descriptive, hermeneutic, and explanatory purposes, but our focus here on their use for describing and displaying qualitative data.

Miles, Huberman, and Saldana (2014) outlined a range of displays, with the most common being identified as either a matrix (table) or network and relational diagrams.

Matrices are formed by intersecting variables or labels of interest and presenting these in a table. The evaluator identifies the dimensions for comparison and then uses the data to populate the cells. Cells in the matrix can be raw data or summaries of statements. Matrices may allow comparisons across evaluation participants or sites, and can be used to organize and display data by type (that is, interviews, observations). For example, in a role-ordered matrix, the row headings may be the stakeholder role (nurse, surgeon, social worker) with the columns indicating responses to structured or semi-structured interview questions. In a time-ordered matrix, the evaluator may classify data according to the program timeframe to show change or progress over time in several sites implementing a new program. It is clear that some form of coding and categorizing is required to determine the content of the cells.

Network models or diagrams display and organize observations, data, or events in relationship to each other. Hierarchical tree diagrams can depict relationships among particular concepts or dimensions. These types of diagrams emphasize the power of visual displays in contributing to understanding. When included in evaluation reports, they assist readers in making the links between the evidence collected and key claims.

Steps to Develop Matrix Displays

1. Clarify the purpose for using matrix displays or network diagrams.
2. Assess the data types and sources available and consider the way these data types and sources can be displayed. Matrices can be developed during data collection in the form of templates that can be filled in by the evaluation team, or be developed post–data collection to display and summarize key findings.
3. Identify categorization and classification options. Identify the type of matrix that will be most useful given the evaluation context and the key evaluation questions. The evaluator may initially develop matrices that summarize findings from different data types (interview or focus groups) or sources (sites or participants). Frequency/occurrence matrices and binary dichotomous matrices for key issues (yes/no; high/low) may also be useful organization schemas.

4. Define the format for display. In general, a landscape presentation of a matrix is easier to read than a portrait version. Matrices work best with a readable number of rows and columns, ideally fewer than twenty. If the evaluation includes a range of sites or a large number of people, you will have to identify ways to cluster and reduce the dimensions further, for example, by type of school (primary schools and secondary schools, rather than individual schools) or role of informant (director, manager, employee, rather than individual). In this way, data from a larger range of stakeholders can be represented in the matrix. Select column and row labels to allow comparison within and between cells and begin filling in the cells. Gaps in cells may indicate missing data if the summary was based on structured or standardized interviews.
5. Compare and contrast both within columns and across rows to form propositions. In this step the matrix is used to inform description and to explore relationships between dimensions.
6. Use a range of matrices to support description and analysis. Often more than one matrix will be appropriate, particularly if there is an interest in more layered analysis.
7. Present findings with reference to the matrices or diagrams.

The evaluator may elect to include one or two matrices in the report to provide evidence for evaluation claims, or to illustrate the categorization process. Some evaluators use matrices to make the process of description, analysis, and interpretation transparent: the reader can trace the interpretations back to the descriptive matrices included in the report (or in a technical appendix) that accompanies the report. This may be particularly important when findings are controversial or there are mixed stakeholder perspectives about the value of the program. Avoid the use of too many diagrams or tables in final reports, as they may detract from key messages and overwhelm the reader.

Application

An example of a role-ordered matrix is presented in Table 22.3. This matrix was presented at the end of a narrative section as a qualitative summary of key points and issues raised by stakeholders about a quality improvement program for breast disease services. The columns indicate the dimensions summarized in the matrix. The rows represent a summary of the perspectives shared by various stakeholder groups.

TABLE 22.3. SUMMARY MATRIX: PERSPECTIVES AND EXAMPLES OF QUOTES.

Stakeholder Group	Perspective on the Program	Example Quotes
Consumers	Favorable about the program, but unsure about evidence of change Questioned the expenditure on funding, especially non-tangibles Unsure about others' acknowledgement of their role in direction of the program Tended to highlight and profile visible (tangible) outputs of the program (e.g., the hand-held record)	"It's a really vital thing to do and it will improve services, especially for rural women." "Our involvement [in the program] is often tokenistic." "There needs to be a broader consumer base to draw on. I feel I am the token consumer voice."
Radiologists/ Oncologists	Saw value in the initiatives undertaken in the program, but concerned about the inequity of service delivery to other cancer groups	"I see a lot of cancer patients, and every one of my cancer patients has the same problem. It's just happens that one of them . . . the breast cancer group . . . gets the golden run." "The rest of the cancer community may be affronted if more money is given to breast cancer and the rest are ignored."
Specialist Breast Surgeons/ Breast Care Nurses/ Nursing Managers/ Clinical Coordinators	Saw value in the program Many stakeholders across these groups highlighted specific changes in practice they attributed to the program Concern about sustainability and transferability of the model to other cancers	"The institution sees breast cancer as a growth area, as a high profile area that is worthwhile to be involved in. They see it as a program that can be generalized to other cancers." "I see it as a model that could be readily transferred to other cancer groups."

When These Methods Are Appropriate

Matrices are commonly used in evaluation as a way to bring together salient information from across cases or across data types in one format. During the analysis process, they provide a structured, visual display of data that allows the evaluator to compare and contrast information, identify patterns, and explore outliers. In evaluation, the progressive use of matrix displays will lead to propositions that can be tested in further stages of data collection.

In written reports, we have found that evaluation audiences appreciate the inclusion of a summary table and may often prefer this form of representation to a lot of freestanding text or narrative description. They want to quickly grasp the key points.

Evaluators who adopt descriptive methods are often driven by pragmatic concerns about coding and classification. Classifications that make sense to the evaluation purpose and for presentation enhance their credibility to evaluation audiences.

Hermeneutic Methods

Overview. Hermeneutic methods often involve an iterative process between data collection and analysis and acknowledge the role of the evaluator and their observations as part of the data corpus. These approaches attempt to generate manifest and latent meanings inherent in qualitative data by some process of coding and labeling. Manifest meanings are visible, descriptive labels that appear in the data. Latent meanings are underlying meanings, gleaned from an iterative process of examining the material, looking for similarities and differences, and identifying themes.

Some hermeneutic methods, such as narrative analysis, eschew formal coding in favor of case-based knowledge and holistic accounts. Evaluators from these traditions contend that the process of coding fragments the data and strips meaning from the context in which it was produced.

Most often, evaluators who utilize qualitative methods will draw on hermeneutic processes, in recognition of the value of rich data and the role of the evaluator in generating interpretations and judgments from the material. The processes associated with textual analysis in the qualitative evaluation literature also tends to emphasize these approaches. The popularity of hermeneutic methods reflects the paradigmatic commitments of many evaluators, influenced by interpretive frames of reference that emphasize a nuanced, contextualized response to textual and visual materials.

Example of Hermeneutic Method: Thematic Analysis. A common type of analysis associated with hermeneutic methods is thematic analysis. One of the features of thematic analysis is iterations of analysis over time. The evaluator moves back and forth among data, classification, and writing about data. The first stage of analysis is often used to inform further data collection. In this case, the evaluator will return to the field to gather more data.

Thematic analysis can be deductively or inductively driven, that is, the evaluator can assign data to categories from existing program theory or according to key evaluation questions. More commonly, thematic data analysis is associated with inductive reasoning, in which the evaluator reviews the material and generates organizing categories that adequately summarize the content. In practice, we believe that inductive, deductive, and retroductive reasoning are at play in analysis of materials generated within an evaluation context. Broadly speaking, induction entails using observations from data gathering to infer to a general conclusion. Deduction involves developing a hypothesized statement and looking for evidence in the study.

In retroduction, conclusions or claims are made and then an assiduous search for confirming and disconfirming evidence is undertaken. The reasoning is based on trying to answer the question, "How did this come about?" While retroduction sounds similar to deduction, the conclusion must follow from the premises.

The evaluator has collected the information from participants to answer particular key evaluation questions. These orient the categorization and classification of data. While the evaluator should remain open to categories inherent in the data, more often than not a combination of strategies is required.

Thematic analysis broadly involves the following eight steps:

1. Clarify the purpose for using thematic analysis.
2. Focus the analysis process by revisiting the key evaluation questions that framed data collection.
3. Familiarize yourself with the materials to be analyzed. The task in this first step is to organize existing material. The material may include transcripts from individual or focus group interviews, documents, observational notes and evaluation reflections, and/or materials produced by the participants (diaries, drawings). The evaluator may have a partial data set that will be analyzed to inform further stages of data collection, or may have a complete data set.
4. Undertake first-level or open coding. First-level coding involves reviewing the material and assigning labels or tags to text, video, or observation segments. The purpose is to descriptively capture what the individual or group is talking about or what is occurring. The coding process, whether formal or informal, is a data reduction process. Its purpose is to tag text or material in order to allow classification of data into a smaller, more refined set of labels. Coding is a common strategy in developing broader category clusters.
5. Do second-level/pattern coding and categorizing of the material. This step involves looking for patterns within the material and across other materials

that have been coded. Similarities and differences are explored and these are used to assign categories. At this stage, the evaluator may have up to thirty or more categories. Categories can also be identified from program theory and used to classify evaluation materials, but the evaluator should ensure that they remain open to other categories evident in the data.

6. Write a memo about relationships and connections derived from the coding and categorization process. Writing or diagramming assists the evaluator in forming propositions. The evaluator may also explore the connectedness among data elements, even when these elements are not directly comparable in terms of similarities or differences. Memos are a powerful strategy that draws on the interpretive skills of the evaluator.
7. Elaborate a limited set of themes. Identify both manifest and latent themes with example quotes. A manifest theme may be the reference to an experience that the majority or all of the participants discussed. For example, most participants may have expressed fear and anxiety about participating in the new program. A latent theme is generated by the evaluator and is often developed from exploration of a broader corpus of data. The evaluator may search the transcript for particular word choices used by the participants, level of emotion expressed, comments that were emphasized in some way (for example, the most important thing you need to know about this program is that . . .) sideline or off the record comments, unexpected or unanticipated comments that surprised the interviewer/evaluator, or things that were not said during the interview. The evaluator's observations during site visits or during the data collection process may contribute to identification of latent themes. A careful audit trail must be maintained to ensure the process of developing themes is transparent.
8. Begin writing about categories and their relationship to codes and emerging insights from the analysis process. While this is defined here as the final step, in practice, writing is a way of finding out (Richardson, 2007) and occurs throughout the process, during stages of writing memos and in drafting of reports. Test propositions by referring back to the data. Diagram relationships using matrix displays and network diagrams or mind maps, and identify example quotes to illustrate the key claims about the program.

Other hermeneutic forms do not privilege coding as described above. For example, narrative approaches in evaluation often seek to retain the sequential and storied elements of an interview and eschew the fragmentation that occurs with formal coding and classification. In narrative analysis, some kind of categorization process is evident, but formal coding may not be. Clearly, the analyst is making decisions about which elements (or categories of meaning) to present.

Seidman (2005) noted the importance of creating profiles from interviews to maintain the coherence of the account from the participants' perspective. Michael Patton (2015) has also written about the value of creating case profiles to make sense of data.

Application

A thematic analysis was adopted to understand the experiences of early years educators involved in a coaching program. The evaluation included a combination of qualitative (interviews and focus groups) and quantitative methods (a statewide survey).

Full transcripts of interviews with coaches and educators were prepared and used as the basis of analysis. Transcripts were sorted by perspective (coaches and educators) about the contribution of the program to observed outcomes. Transcripts from each group were read and notations made in the right-hand margin in the form of categories relating to their experience and to the key evaluation questions. For example, positive views of the program, negative views of the program. Illustrative quotes that exemplified the categories were highlighted. In the left margin, the evaluator made comments about issues to follow up or other evaluation information (links between survey information and interviews). Memos about each transcript were made, and a collation of categories and related codes was recorded. A within-case analysis was undertaken (within each stakeholder group) and a summary made of the key findings associated with key evaluation questions. A cross-case analysis (of comments made by coaches and educators) was then undertaken to identify salient program variables that may have contributed to outcomes. Illustrative quotes provided evidence for evaluation claims, and this information was combined with the available quantitative data (a statewide survey) to answer evaluation questions.

When These Methods Are Appropriate

Hermeneutic approaches are valuable if the evaluator has access to rich data in the source of interview transcripts or comprehensive notes of observations at site visits. But the downside is that hermeneutic methods often require the investment of more time than enumerative or descriptive methods. The evaluator needs to organize the data, identify codes and categories, and compare and contrast segments of the data to identify patterns and themes, often returning to the data to generate and test these propositions.

One of the key features of hermeneutic methods is the maintenance of the depth and detail of the data collected. There is likely to be fuller coverage of the available data, rather than selective coding of pieces or sections deemed to be most closely tied to the evaluation questions.

In larger studies, time may prohibit the inductive identification of codes and categories. Under these circumstances, the evaluator may decide to inductively code and categorize a proportion of the materials and then apply this coding and categorizing system to the remaining set. The evaluator should still retain an openness to new or emergent codes that are not captured by the existing classification system.

Those evaluators that use hermeneutic methods often privilege constructivist criteria such as trustworthiness and authenticity, rather than traditional validity (truth value) criteria. Specific strategies, such as reflexivity, triangulation, and respondent validation may be adopted. Those evaluators that hold a critical or social change perspective, may focus on a more explicit set of relational and ethical criteria (Abma and Widdershoven, 2011; Lather, 1993) that involve an assessment of the consequences of the findings, the capacity of those involved to take meaningful action, and the explication of values across a range of stakeholder groups.

Explanatory Methods

Overview. Explanatory methods of analysis focus on generating and testing causal claims from data and are particularly relevant to address outcome evaluation questions. In impact evaluations, clients want to know whether the program worked, and if so, how it worked to produce outcomes and impacts.

Most evaluation clients want to understand the contribution of projects to the larger program of effort, or they seek transferable lessons from implementation of projects in an array of contexts. Others want to understand what programmatic factors in what combination contributed to the observed outcomes so they can make decisions about future funding. For example, do the support resources and professional development workshops together foster improved knowledge and skills of welfare workers, or would the professional development workshops be sufficient on their own to generate improvements?

These are causal or explanatory questions. They require analysis that can link proposed causal mechanisms to specific outcomes. In traditional terms, the role of causal analysis has been limited exclusively to the use of experimental or quasi-experimental designs. Yet, most evaluators are called upon to address these questions when these designs are not relevant (for example, quality improvement programs implemented across a state or borough, or when implementation has already occurred). Other techniques will be required to address these questions.

There is growing interest in alternatives to experimentation for strengthening causal inferences, and a number of promising strategies are available utilizing qualitative data in small n studies. Process tracing and qualitative comparative analysis are two potential strategies that may contribute to plausible causal claims. In this section we focus on qualitative comparative analysis as a promising analytic method for addressing explanatory claims in evaluation contexts.

Example of Explanatory Method: Qualitative Comparative Analysis. Qualitative comparative analysis (QCA) is a systematic comparative approach that maintains a focus on the richness and context of cases. Charles Ragin (1987), a political scientist, pioneered the use of the method arguing for its relevance in determining causal claims in small n studies. Between five to fifty cases or more are appropriate for QCA.

Essentially, QCA examines the combinations of factors that are associated with a particular outcome using a minimization algorithm generated from comparing and contrasting causal conditions and their relationships to an outcome of interest.

There are two types of QCA: crisp set qca and fuzzy set qca. In crisp set qca, each individual condition identified for each case (school, program, and so forth) is classified with an 0 or 1 to indicate absence or presence, respectively, of this condition. Fuzzy set qca allows levels of presence or absence (for example, .25, .50, .75).

Broadly, the QCA method involves the following steps:

1. Clarify the purpose and appropriateness of using qualitative comparative analysis
2. Identify the outcome of interest and the relevant conditions that may be associated with the outcome. This is a critical stage and must be strongly informed by theory. The evaluator can draw upon program theory or policy documentation or consultations with stakeholders to identify a limited number of conditions.
3. Develop a truth table and classify cases as rows and conditions as columns. A truth table is a matrix that includes various combinations of conditions and the outcome. Each identified causal condition is categorized in the truth table by its presence or absence (crisp set) or by the level of presence

or absence (fuzzy set) in relation to the outcome. We acknowledge that there may be some loss of data precision in this step, but the formalization of decision rules about classification is helpful in generating and testing propositions. Criteria for classification must be transparent. This process also enables to evaluator to examine patterns among causal conditions and their relationship to the outcome of interest. An example of a truth table is presented in Table 22.4.

4. Using the software of choice (for example, NVivo [QSR)] Tosmana [Cronqvist, 2006]), compare the constellations and interactions among conditions and the outcome.

The output provides the minimization algorithm, using Boolean logic (or + / and x) that identifies the specific combination of variables that are sufficient for occurrence or non-occurrence of the outcome. The notion of equifinality is important here. QCA is based on a conception that it is possible that an outcome can be achieved from different patterns of causal conditions and that a causal condition (for example, clarity of partnership roles) may have a different influence on an outcome, depending on the context. The evaluator can explore which conditions are necessary for outcome success. While this method appears to have much in common with regression analyses, there is a substantive difference in underlying philosophy. QCA depends on deep case knowledge to generate causal claims and is based on analyses of set-theoretic relationships, not variable-oriented relationships.

5. Explore contradictory configurations for cases to generate a richer knowledge of the cases and their contexts. The evaluator may explore the cases involved in the contradictory configuration. Decisions about exclusion of these contradictions will need to be made and be justified clearly in the evaluation report. Interpret the minimal formulas to develop a comprehensive narrative about the causal conditions that contributed to program outcomes in particular contexts. Use case-based knowledge to ground propositions about the relationship between particular conditions and the outcomes of interest.

Application

Befani, Ledermann, and Sager (2007), used QCA in an evaluation of the effects of the Swiss Environmental Impact Assessment (EIA). The EIA is a regulatory structure that ensures infrastructure projects address environmental

TABLE 22.4. SIMPLE EXAMPLE OF A TRUTH TABLE.

Case	Clarity of Partner-ship Roles	Grants (High Grant Success)	Registration High	Broker Stable Service of Broker)	Community Plan in Place	Community Development Activity High	Outcome (Success)
1 Urban	1	1	1	0	1	1	1
2 Urban	1	0	0	0	1	1	0
3 Rural	1	0	1	1	1	1	1
4 Rural	1	1	0	1	0	0	0
5 Urban	0	0	1	1	0	0	0
6 Urban	0	0	1	0	1	0	0
7 Rural	1	1	0	1	1	1	1

standards. After reviewing theory and existing studies, the team identified plausible configurations of conditions that may trigger outcomes within particular contexts. In essence they developed a program theory about the ways in which an outcome of success or failure may occur. Conditions were assigned a value of 1 if they were present and 0 if they were absent. Case studies were then undertaken to ensure thorough knowledge of the cases, and to generate a truth table of conditions and trajectories of outcomes (failure or success). QCA was then performed with the use of fuzzy set QCA (fs/QCA). The output from these analyses led the authors to a series of propositions about the influence of the policy context on outcomes. The authors argued that the application of QCA was consistent with their adoption of a realist evaluation framework and that the process involved necessary iterations among theory, data, and propositions.

When These Methods Are Appropriate

Explanatory methods may be valuable in impact evaluations. Qualitative comparative analysis holds much promise for evaluators who seek to maintain a close and rich understanding of the context of particular cases, but are also interested in testable patterns within and across cases. Given the central role of the truth table and the minimization formula, it is important to use software that supports these processes. Specific software developed for QCA may be used (for example, Tosmana) or generic theory testing software such as NVivo can be used to organize and classify the data.

Evaluators who adopt explanatory approaches such as QCA will tend to draw on traditional scientific criteria, but while traditional canons of reliability and validity are held, they are reinscribed in meaning. Advocates of QCA adapt the traditional criteria to reflect the importance of case knowledge, emergence and iteration, and interrelationships between observed outcomes and configurations of conditions that are associated with those outcomes. Given the inherent interest in causal mechanisms and context, QCA will be relevant to evaluators who practice realist evaluation

Framing Analytic Choices

We have presented a typology of four purposes of qualitative data analysis and focused in on one method that is closely associated with the purpose, enumerative, descriptive, hermeneutic, and explanatory. We have argued that it is better to choose data analysis methods that fit the purpose of the analysis, the context of the evaluation, and the technical requirements of the process rather than privilege one method. Evaluators have a range of analysis options that can be used and, ideally, using more than one method (analytic triangulation) may strengthen the plausibility and persuasiveness of evaluation claims.

Choice of analytic method will be influenced by the nature of the data, purpose, and desired product. For some audiences, a numeric summary of the material gathered by frequency may be preferred. For other audiences, the use of rich illustrative quotes will be favored. Or the evaluator may choose a combination of strategies to portray the findings and provide warrants for key claims.

Some analysis and presentational strategies require particular kinds of data. For example, there may be limited value in applying narrative methods of analysis if the evaluation has only generated short, very structured interview material. Narrative approaches often utilize free-flowing conversation, rather than information presented in a structured question-and-answer format. Similarly, it may not be appropriate to quantify material that was designed to illuminate or illustrate an issue or to use semi-structured or open-guide interviews when individuals are not asked to share insights on a particular issue. We are not implying that the evaluator cannot transform this material into another type of product, but caution that the evaluator be aware of strengths and limitations of particular techniques according to available data.

Potential data analysis options should be anticipated before any data are collected, to ensure that there is alignment between type of information required, scope of data, and purpose. As analysis of qualitative data is often undertaken concurrently with data collection, the evaluator can rehearse data analysis and display options to maximize communicative power.

How Can Software Help?

Many different computer packages facilitate the coding and analysis process. These can be categorized into four groups:

- Word processing and spreadsheet software, such as Word and Excel, which can be used for simple coding and analysis
- Specialized qualitative data analysis software, such as NVivo, HyperResearch, MaxQDA, and Atlas-TI
- Machine learning software, such as MonkeyLearn
- Visual analysis software that produces word cloud or network diagrams of text

Word processing and spreadsheet software can be useful for simple analysis of small quantities of data where there is not the time or budget to buy and learn specialist software. Text (such as the answers to an open-ended question) is put into cells, one cell per response. A label is given to each of these and it is typed into an adjoining cell. The data can then be sorted and the coding checked for consistency. Using a spreadsheet such as Excel makes it possible to use pivot tables to produce summary statistics, such as counts for each code and cross-tabulations with other variables, such as gender.

Specialist qualitative data analysis software do not automatically analyze the data, but provide structures for organizing, classifying, and storing information. They can do some autocoding, for example, doing a text search and creating a category of all text that includes that code, or putting all the responses to a particular questionnaire item or interview question into a category for further analysis. Many packages allow versions of coding to be stored for later retrieval, which can be helpful in tracking movement in the sophistication of the analysis and allow documentation that justifies the links among data, evidence, and evaluation claims.

Specialist software is recommended where this is a large volume of data and analysis (for example, multiple stakeholder groups, or thirty or more cases), and where there is the time and money to buy and learn how to use it. Software facilitates the efficiency and tracking of the coding and retrieval process. The most common error in using this type of software is to become stuck at the coding stage, creating and using an impractical number of codes. For evaluations, in most cases it is more useful to do an initial cycle of coding, analysis, and reporting and then return to the data with more questions and do further coding. This reduces the risk of not producing a useful report.

Machine learning software is designed to learn how to classify data. For example, a sample of hotel reviews can be classified as either "positive" or "negative." The software will identify predictive words and phrases (for example, "disgusting" or "worst hotel") and show the degree of success in predicting the classification based on these words. If the level is not sufficiently high, further training with more data can be undertaken. Once an acceptable level has been reached, the software can be used to automatically code data. This type of software is most appropriate when there is a large quantity of data (for example, using a web scraper to take data from reviews or social media) and a simple classification needed (for example, positive or negative).

Visual analysis software can produce a word cloud, where the size of a word reflects its frequency, and word network diagrams, which show how words are connected in the text. While word clouds have become increasingly popular as a quick way of sharing some analysis, they are often not very helpful. For example, doing a word cloud of book titles, speeches, or newspaper reports will have all the risks of simple enumeration previously discussed. Even if words such as "the" and "and" are excluded, word clouds don't deal with synonyms well. They can be useful when the data have been collected in the form of a single word summary.

Software developments are ongoing and rapid. The CAQDAS website (http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/index.htm) offers regular updates on software developments. While multiple texts on the attributes and use of software are available, many of these do not frame the discussion of software in relation to qualitative research and evaluation practices. Two notable exceptions are Kurkkartz (2014) and Richards (2009). These texts discuss analytic practices and profile the use of specific packages, MaxQDA and NVivo, respectively.

Given the range of software available, St. John and Johnson (2000, p. 397) have identified questions that researchers or evaluators might ask themselves when choosing a package, including the following:

- "What are the advantages and disadvantages of this package for my research?
- What purpose will this package serve for this research project?
- Will this package handle the type of data I intend to collect?
- Will this package enable flexible handling of data for this project? …
- Will this package enable me to interact with and reduce data in a way that is consistent with my methods?"

Who Does the Analysis?

In larger evaluations, we may be working as part of a team. Team members may be assigned different tasks such as data collection or instrument design across all sites, or be responsible for collating and analyzing all information from a project site with the expectation that the information will inform cross-site analysis. There are distinct advantages of analyzing with others, most notably the presence of critical friends to challenge and test assumptions and claims and as a form of triangulation in the analysis process. However, if there is a lack of clarity about processes of analysis and how the team will discuss coding and categorizing processes, the resulting product can be very messy. We recommend that teams discuss analysis options and clarify processes that will be adopted prior to data collection. Review and discuss initial coding with team members of a small subset of the data before proceeding to analyze the whole data set.

Co-analysis can also be undertaken with clients as a form of early data rehearsal or as a way to identify implications of the data. This may be particularly fruitful when the evaluator does not have substantive knowledge base about the evaluand. Good interpretations occur within a framework of context and descriptive detail. For example, Del undertook an evaluation of a statewide coaching program (2013) for early year professionals. She did not have a background or substantive experience with early years services. Initial propositions were discussed with a reference group, including experienced early years coaches and program funders who had considerable experience in early education and care. Interpretations were reviewed by drawing on the evidence collected in conjunction with the substantive content knowledge base shared by the coaches and funders.

High Quality Qualitative Data Analysis

We have argued that methods of analysis should reflect the purpose of the analysis and the intended product required. Good quality analysis of qualitative data requires clarity about which standards will be used to judge the quality and effective strategies to meet these standards.

In this section we focus particularly on the issue of the quality of inferences—a reframing of the standard of accuracy to be more appropriate across the different approaches to qualitative data analysis. Traditionally, the terms “validity” or “trustworthiness” have been used to denote an empiricist or interpretivist orientation to evaluation with an associated series of procedures that evaluators may adopt. The mixed method literature has moved toward the term “inference quality.” Rather than remain tangled in the debates about the most appropriate terminology for quality and given that most evaluations adopt multiple methods, this term seems appropriate as an umbrella term.

Depending on the paradigmatic stance of the evaluator and the intended users of the evaluation, there can be quite different standards for what is considered to be high-quality inference in qualitative data analysis. Constructivist evaluators may highlight the value of member checking or respondent validation and triangulation. Evaluators who adopt a critical change focus to guide their work (Mertens, 2009; Wadsworth, 2011), may draw more on strategies that engage with issues of power and diversity and advocate relational criteria such as authenticity and reciprocity.

Program Evaluation Standards and Quality criteria for QDA

The five high-level evaluation standards—utility, accuracy, feasibility, propriety, and accountability—provide a useful reference point for thinking about the quality of qualitative data analysis. While the accuracy standard seems most closely associated with quality of data analysis, the standards broaden consideration of analysis as purely a technical activity.

Utility is about usefulness and fit for purpose. Analysis methods must be selected that fit the purpose and intended product required. In some cases, we may select matrix displays to guide analysis and use these in a written report to substantiate the claims we make. Matrix displays can summarize the evidence in a format that is readily understandable to the client. Evaluators may combine criteria or choose dimensions that reflect the intended purpose of the analysis.

The second standard is *accuracy*. Accuracy applied to qualitative analysis means that the evaluator will pay attention to issues of validity and trustworthiness of inferences generated from the analysis process. The considerations of analytic rigor will differ depending on analytic method. For example, inter-rater reliability may be a preferred strategy in classical content analysis to strengthen rigor, but it may be rejected as a concept of relevance to a constructivist or critical/transformative evaluator. For example, interpretive differences between evaluators do not mean that one evaluator is wrong, just as a high inter-rater reliability coefficient does not confer correctness or accuracy.

A number of strategies are adopted by evaluators to strengthen inferences from qualitative data. In descriptive and hermeneutic methods of analysis, evaluators are likely to draw from the trustworthiness criteria developed by Lincoln and Guba (1985) A transformative evaluator may find the authenticity criteria associated with collaborative and participatory approaches more relevant as a guide to strengthen inference quality. Those evaluators adopting explanatory methods will likely be influenced by criteria such as pattern matching and assessing alternative explanations.

While there is some overlap among the strategies for supporting accuracy of inferences and acknowledging the interpretive basis of qualitative data analysis, the evaluator must be able to defend decisions. Evaluators may combine criteria or choose dimensions that reflect the intended purpose of the analysis, but it will be important to maintain consistency between the techniques adopted and the evaluators' perspective. For example, many advocates of thematic analysis undertaken from a hermeneutic perspective would eschew the calculation of inter-rater reliability, especially if the interviews were unstructured or narrative accounts. Inter-rater reliability confers confidence in the reader that two or more analysts have coded consistently, which communicates a message to evaluation audiences that the higher level of agreement means that the coding is not only consistent, but right. For many interpretivist evaluators, inter-rater reliability is a superficial technique that does not add value and does not recognize the inherent intersubjective nature of qualitative methods. This does not presume that all classifications are equally valid, but rather challenges the conception that reliability is a precursor to accuracy.

The interpretive strategies outlined below in italics are not designed to be used as a checklist, as some will be more relevant than others in differing evaluation contexts and for differing analytic purposes.

Trustworthiness strategies include activities that are concerned with the depth of evidence gathered such as *prolonged engagement.* In an evaluation context, this strategy may lead evaluators to question whether their exposure to the program and to data collection opportunities provided a sufficient basis to make particular claims about implementation and outcomes. A second strategy is *triangulation.* Evaluators may be more confident in the credibility of their claims if there is corroboration of evidence across data sources, methods, theories, or analysis processes. It is important, however, to be aware that divergent evidence may also be useful in encouraging the evaluator to test assumptions and to more deeply consider alternate explanations. *Peer debriefing* with other team members, evaluation advisory groups, and those not directly involved in the evaluation can help the evaluator avoid blind spots and can assist the evaluator to make defensible claims.

A fourth strategy is to review *negative cases.* The evaluator pays attention to cases that do not support or, indeed, contradict patterns in the data. In the evaluation of a training program, the evaluator may seek out those participants who did not complete the program, rather than relying only on those participants who do complete. The evaluator may return to participants to discuss interpretations through a *member checking* or *respondent validation* process. Participants may be asked to review interpretations and/or provide additional information to sharpen interpretations. In some collaborative evaluation approaches, this process is explicitly built into the evaluation from the planning stages. Methodological accountability is supported through strategies such as reflexivity and audit processes. *Reflexivity* means that evaluators critically assess their role in data collection and analysis, their potential influence on participants, and the values that have shaped their interpretations.

An *audit trail* is a useful strategy to strengthen the rigor of claims. Evaluators have to document their data collection processes and decisions and ensure transparency in linkages among data collection, analysis, and evaluation claims so that these can be reviewed and assessed by others. Evaluators adopting participatory or transformative evaluation approaches are likely to draw on trustworthiness strategies, but supplement these with attention to authenticity criteria that acknowledge issues of participant access, voice, power, and representation.

Feasibility is an evaluation standard that relates to prudence of processes and procedures. In a data analysis context, this means that methods should be selected that are appropriate to the budget and timeframe available for the evaluation. Hermeneutic and explanatory methods are likely to take much longer and entail more iterations than enumerative or descriptive options. The evaluator must consider the feasibility of methods at several stages of the evaluation, balancing attention to information needs, rigor, and timelines.

Propriety in the evaluation standards refers to the need for evaluations to be ethical and legally appropriate. The depth and rich detail of qualitative data means that the evaluator must carefully consider how best to protect the identity of participants and respect their rights within the evaluation and in the written presentation of the evaluation report. The evaluator must have appropriate processes in place to ensure raw data files are appropriately stored and that pseudonyms, if appropriate, are applied. Inessential details within the context of the evaluation (such as number of siblings or age) can be modified to further protect identification. If modifications are made in this way to protect participants, this should be declared in the method section of the evaluation report.

Accountability is the fifth evaluation standard. In program evaluation, this is most closely associated with meta-evaluation. In the context of data analysis, accountability means that the evaluator documents purposes, decision rules, and processes and ensures that a clear argument is made for the selection of particular analytic approaches. In the spirit of transparency and accountability, the evaluator should ensure that materials that form the basis of the analysis are available for secondary review if required, as long as appropriate ethical issues regarding access to primary data have been addressed.

The standards can sometimes be in tension, so consideration has to be given to the appropriate level of trade-offs. For example, achieving a high standard of accuracy is constrained by the need to be timely (utility standard), cost-effective (feasibility standard),and not excessively burdensome on respondents (propriety standard).

Table 22.5 summarizes the application of the evaluation standards to qualitative data analysis processes. The strategies are suitable for all methods of analysis discussed in this chapter, with particular strategies associated with particular methods where relevant.

TABLE 22.5. EVALUATION STANDARDS AND STRATEGIES TO MAXIMIZE INFERENCE QUALITY.

Evaluation Standards	Qualitative Data Analysis Principles	Strategies to Consider
Utility	Data analysis decisions should be fit for purpose and relevant to client needs	Potential data analysis methods should be identified before data collection begins and then reviewed throughout the evaluation Requirements for each type of analysis method should be made explicit Strengths and limitations of analytic methods should inform selection
Accuracy	The analysis process should be rigorous and transparent according to method	EnumerativeInter-rater reliability, documentation Descriptive/hermeneutic: Use relevant strategies to support inferences such as peer debriefing, member checking, negative case analysis, reflexivity, triangulation, audit trail Hermeneutic/explanatory Pattern matching, ruling out alternative explanations
Feasibility	The choice of analysis method should be feasible given purpose, resources, and timeframes	Identify data requirements for analysis purpose and build in time/cost to evaluation budget
Propriety	Analysis processes should protect the rights of participants	Discuss limits of confidentiality with participants and client Define likely risks of analysis processes
Accountability	Analysis processes should be documented and linkages among data, analysis, interpretation, and reporting should be explicit	Agree on procedures for secondary review/peer review of analysis

Conclusion

Qualitative data make a powerful contribution to research and program evaluation. Techniques for the effective gathering of such data are important, and many texts detail these processes well. Similarly, methods of data analysis have been described, but have not themselves been amenable to analysis for their suitability for purpose. This has often resulted in analyses that are overly shallow or descriptive analyses that undermine the value of the information generated.

In this chapter, we have argued that decisions about analytic techniques be considered early in an evaluation, and that these decisions will be informed by a range of practical considerations, such as audience, time and resource constraints, and the type of data available. Whatever approach or range of approaches are adopted to code or summarize, classify, analyze, and synthesize qualitative data, there is a need to make sure the processes are transparent and traceable. This is not for the purpose of replicability, but for the justification of claims that arise from purpose of the analysis.

We have presented a typology of purposes and outlined an example of a key analysis method for each purpose. The typology is presented as a heuristic, not intended to be prescriptive, as this would negate the value of deliberative decisions based on an appreciation of the specific context. Evaluators may combine purposes and methods in a single evaluation or mix analytic methods to challenge existing insights or to generate new ideas.

An overview of quality criteria for assessing and strengthening data analysis was provided that builds on substantive guidance provided by the program evaluation standards. The warrants for claims evaluators make must be based on strong, clear synthesis of evidence. Evaluators require a rigorous set of methods for making sense of qualitative data and for assessing its appropriateness. This chapter has reinforced the value of systematic analysis, creativity, and the important of transparency of these processes to the evaluation profession and to evaluation clients.

A quote by Janice Morse (1994) provides a fitting conclusion that captures the technical and creative elements of good qualitative data analysis:

> Data analysis is a process that requires astute questioning, a relentless search for answers, active observation, and accurate recall. It is a process of piecing together data, of making the invisible obvious, of recognizing the significant from the insignificant, of linking seemingly unrelated facts logically, of fitting categories one with another, and of attributing consequences to antecedents. It is a process of conjecture and verification, of correction and modification, of suggestion and defense. It is a creative process of organizing data so that the analytic scheme will appear obvious. (p. 25)

References

Abma, T., and Widdershoven, G.A.M. "Evaluation as a Relationally Responsive Practice." In N. Denzin and Y. Lincoln (eds.), *Handbook of Qualitative Research* (4th ed.). Thousand Oaks, CA: Sage, 2011.

Baptiste, Ian. "Qualitative Data Analysis: Common Phases, Strategic Differences." *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* [Online Journal], September 2001, *2*(3). Available at www.qualitative-research.net/fqs-texte/3-01/3-01baptiste-e.htm

Bazeley, P. *Qualitative Data Analysis: Practical Strategies*. London: Sage, 2013.

Befani, B., Ledermann, S., and Sager, F. "Realistic Evaluation and QCA Conceptual Parallels and an Empirical Application." *Evaluation*, 2007, *13*(2), 171–192.

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

Claburn, T. "How Google Flu Trends Blew It." *InformationWeek,* September 25, 2013.

Collier, D. "Understanding Process Tracing." *Political Science and Politics,* 2011, *44*(4), 823–830.

Crabtree, B. F., and Miller, W. L. *Doing Qualitative Research*. Thousand Oaks, CA: Sage, 1999.

Cronqvist, L. *Tosmana–Tool for Small N Analysis* (version 1.254). Marburg, Germany: University of Marburg, 2006.

Gladwin, C. H. *Ethnographic Decision Tree Modeling*. Thousand Oaks, CA: Sage, 1989.

Kuckkartz, U. *Qualitative Text Analysis: A Guide to Methods, Practice, and Using Software*. Thousand Oaks, CA: Sage, 2014.

Krippendorf, K. *Content Analysis: An Introduction to the Methodology* (3rd ed.). Thousand Oaks, CA: Sage, 2013.

Lather, Patti. "Fertile Obsession: Validity After Poststructuralism." *The Sociological Quarterly*, 1993, *34*(4), 673–693.

Lincoln, Y., and Guba, E. *Naturalistic Inquiry*. Thousand Oaks, CA: Sage, 1985.

Mertens, D. *Transformative Research and Evaluation*. New York: Guilford Press, 2009.

Miles, M. B., Huberman, A. M., and Saldana, J. *Qualitative Data Analysis: A Sourcebook of New Methods* (3rd ed.). Thousand Oaks, CA: Sage, 2014.

Morse, J. "Emerging from the Data: The Cognitive Processes of Analysis in Qualitative Inquiry." In J. M. Morse (ed.), *Critical Issues in Qualitative Research Methods (22–43)*. Thousand Oaks, CA: Sage, 1994

Patton, M. Q. (2015). *Qualitative Research and Evaluation Methods* (4th ed.). Thousand Oaks, CA: Sage, 2015.

Paul, M. J., and Dredze, M. "You Are What You Tweet: Analyzing Twitter for Public Health." ICWSM, International Conference on Weblogs and Social Media, 265–272, July 2011.

Ragin, C. *The Comparative Method. Moving Beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press, 1987.

Richards, L. *Handling Qualitative Data: A Practical Guide* (2nd ed.). London: Sage, 2009.

Richardson, Laurel, and St Pierre, E. "Writing: A Method of Inquiry" (959–978). In N. Denzin and Y. Lincoln (eds.), *The Sage Handbook of Qualitative Research* (3rd ed.). Thousand Oaks, CA: Sage: 2005.

Riessman, C. K. *Narrative Methods for the Human Sciences*. Thousand Oaks, CA: Sage, 2008.

Ritchie, J., and Spencer, L. "Qualitative Data Analysis for Applied Policy Research." In A. Bryman and R. G. Burgess (eds.), *Analyzing Qualitative Data* (173–194). London: Routledge, 1994.

Schreier, M. *Qualitative Content Analysis in Practice.* Thousand Oaks, CA: Sage, 2012.

Seidman, I. *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences.* New York: Teachers College Press, Columbia University, 2005.

Semiocast. "Geolocation Analysis of Twitter Accounts and Tweets by Semiocast." http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US, 2012.

Spradley, J. *Participant Observation.* New York: Holt, Rinehart and Winston, 1980.

St. John, W., and Johnson, P. "The Pros and Cons of Data Analysis Software for Qualitative Analysis." *Journal of Nursing Scholarship*, 2000, *32*(4), 393–397.

Strauss, A., and Corbin, J. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* Thousand Oaks, CA: Sage, 1998.

Trochim, W. "Outcome Pattern Matching and Program Theory." *Evaluation and Program Planning*, 1989, *12*(4), 355–366.

UN Global Pulse. *Mining Indonesian Tweets to Understand Food Prices.* New York: UN Global Pulse. www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf, 2014.

Vogt, W. P., Vogt., E. R., Gardner, D. C., and Haeffele, L. M. *Selecting the Right Analyses for Your Data: Quantitative, Qualitative and Mixed Methods.* New York: The Guilford Press, 2014.

Wadsworth, Y. *Everyday Evaluation on the Run: The User-Friendly Introductory Guide to Effective Evaluation* (4th ed.). Sydney, Australia: Allen and Unwin, 2011.

Wetherell, M., Taylor, S., and Yates, S. J. (eds.). *Discourse as Data: A Guide for Analysis.* London: Sage, 2001.