07_184_Ch07.qxd 3/29/07 6:43 AM Page 137

7

Data Reduction Techniques for Large Qualitative Data Sets

EMILY NAMEY, GREG GUEST, LUCY THAIRU, AND LAURA JOHNSON

You have your way. I have my way. As for the right way, the correct way, and the only way, it does not exist.

FRIEDRICH NIETZSCHE

WORKING WITH DATA COLLECTED through a team effort or in multiple sites can be both challenging and rewarding. The sheer size and complexity of the data set sometimes makes the analysis daunting, cbut a large data set may also yield richer and more useful information. In this chapter, we explore strategies for combining qualitative and quantitative analysis techniques for the analysis of large, qualitative data sets. *Large* is a relative term, of course, and many of the techniques described here are applicable to smaller data sets as well. However, the benefits of the data reduction techniques we propose increase as the data sets themselves grow in size and complexity. In our selection of techniques, we have taken a broad view of large qualitative data sets, aiming to highlight trends, relationships, or associations for further analysis, without deemphasizing the importance of the context and richness of the data themselves. This perspective also brings focus to the multiple interpretive lenses that a group of researchers brings to team-based analysis.

Throughout the chapter, we use examples from some of our research to illustrate the use of the methods discussed. In doing so, we identify the strengths and weaknesses of each method and suggest ways in which an appropriate technique may be chosen for a given research question and its corresponding data set.

Approaches to Data Analysis

During the past few decades, qualitative research has greatly benefited from theoretical and methodological developments in data analysis (see, e.g., Bernard and Ryan 1998, Dey 1993, LeCompte and Schensul 1999). Analyses typically fall into one of two categories: content and thematic. In content analysis, the researcher evaluates the frequency and saliency of particular words or phrases in a body of original text data in order to identify keywords or repeated ideas. In addition to simple word counts, content analysis can be expanded to include associated attributes of keywords and other semantic elements, such as synonyms, location in the text, and surrounding words or phrases (Dey 1993:59).

Content analysis techniques are valued for their efficiency and reliability. With appropriate software, large numbers of text files can be quickly scanned and keywords tallied. And since the original, "raw" data are used, there is minimal interpretation involved in the word counts, resulting in greater reliability. The primary drawback to content analysis is that context is usually not considered or is highly constrained, limiting the richness of the summary data produced.

Thematic analysis, in contrast, is more involved and nuanced. Thematic analysis moves beyond counting explicit words or phrases and focuses on identifying and describing both implicit and explicit ideas. Codes developed for ideas or themes are then applied or linked to raw data as summary markers for later analysis, which may include comparing the relative frequencies of themes or topics within a data set, looking for code co-occurrence, or graphically displaying code relationships.

Reliability is of greater concern with thematic analysis than content analysis because research analysts must interpret raw text data in order to apply codes, and because interpretations may vary across analysts. Strategies for monitoring and improving intercoder agreement, and therefore reliability, add to the time required for thematic analysis, but the investment is well worth the context-rich coded data produced (see chapters 6 and 11, this volume).

Both content and thematic analysis can be data-driven, as in grounded theory (Glaser and Strauss 1967, Kearney et al. 1994, Wright 1997), or theory-driven (Krippendorf 1980, Weber 1990). In a data-driven approach, the researcher carefully reads and rereads the data, looking for keywords, trends, themes, or ideas in the data that will help outline the analysis, *before* any analysis takes place. By contrast, a theory-driven approach is guided by specific ideas or hypotheses the researcher wants to assess. The researcher may still closely read the data prior to analysis, but his or her analysis categories have been determined a priori, without consideration of the data.

Theory-driven approaches tend to be more structured, and for this reason may be considered more reliable, in the sense that the same results are likely, regardless of the coder. Conversely, data-driven approaches may be considered to have greater validity because they are more flexible and open to discovery of themes or ideas not previously considered, resulting in theory that is "grounded" in the data. Fortunately, neither approach is so rigid as to prevent borrowing from the other, to maximize the findings of an analysis and to balance reliability and validity. Theory-driven analysis does not preclude the analyst from uncovering emergent, data-driven themes, which may then be added to the analysis, and similarly data-driven analyses may generate theories to explain emergent structure. The techniques presented below span both content and thematic analysis and are applicable to both theory- and data-driven approaches. The examples presented, however, are primarily drawn from data-driven thematic analyses.

Framing the Analysis

Large qualitative data sets generally encompass multiple research questions. Hence, very few, if any, analyses of such data sets simultaneously involve *all* of the data that have been collected. From the outset, researchers need to delineate the boundaries of a given analysis with a comprehensive analysis plan. This plan can include guidelines for data set reduction, including whether all the data will first be coded in an exploratory analysis, whether they will be partitioned in a way appropriate for theoretical analysis and hypothesis testing, or whether some data will simply not be included in specific analyses. Eliminating data not relevant to the analysis at hand—or extracting the data that *are* relevant—is usually the first, and arguably the simplest, form of data reduction. As Miles and Huberman (1994: 11) explain,

> Data reduction is not something separate from analysis. It is *part* of analysis. The researcher's decisions—which data chunks to code and which to pull out, which evolving story to tell—*are all analytic choices*. Data reduction is a form of analysis that sharpens, sorts, focuses, discards, and organizes data in such a way that "final" conclusions can be drawn and verified.

In cases where the larger data set was compiled from more than one type of data collection instrument (e.g., semistructured in-depth interviews, structured focus groups, and pile-sorting activities), the researcher needs to make a decision about the type of data she or he will select from the larger data set. She or he may choose to analyze data from only one type of instrument, or from several different instruments. As described by Patton (1990), this form of "triangulation" across different data collection strategies during data analysis can be particularly helpful when dealing with large data sets. The researcher may also need to frame the analysis in terms of the sources of data or the categories of participants from whom the data were collected (MacQueen and Milstein 1999). This may require limiting the analysis to one or two sites of a multisite project, or limiting the number of subgroups included in the analysis (e.g., including only data relevant in terms of select participant characteristics regardless of site).

The researcher's primary guide in all of these data reduction decisions is a clearly defined analysis objective, accompanied by an analysis plan. The objective may be based on the research questions addressed in the project or it may be determined on the basis of specific reporting or publishing goals. In the analysis plan, the researcher also defines the level of the analysis—whether exploratory, descriptive, hypothesis testing, or comparative in nature—and chooses data accordingly. These initial data reduction decisions are critical to choosing an appropriate strategy for the ensuing analyses.

Analytical Techniques

In this section, we summarize the basic steps, strengths, and weaknesses of several analytical techniques for large qualitative data sets. The techniques do not comprise an exhaustive list; rather, on the basis of our experience in the field, we highlight some of the techniques we have found useful when working as a team in multisite research studies. In addition to these techniques, the fundamental step of reading and rereading text data files is still essential for adding context to theory-driven analyses and for identifying themes for data-driven analyses.

Structural Coding

One effective method for making large qualitative data sets more manageable, for either content or thematic analysis, is to develop and apply a series of "structural" codes to the data. As described by MacQueen et al. in chapter 6 of this volume, the term *structural code* refers to question-based, rather than theme-based, codes. This approach works for data collected using structured or semistructured interview or focus group guides that have discrete questions and probes that are repeated across multiple files in a data set. Each discrete question and its associated probes are assigned a code that is then applied or linked to the question and subsequent response text in each data file. Sets of questions that comprise a conceptual domain of inquiry can also be given a structural code.

Namey and colleagues, for example, worked on a project that included semistructured interviews with participants in an HIV prevention clinical trial. The trial was testing a drug for HIV prevention, and the interviews were designed to gather information on trial participants' experiences throughout the research process. Questions were grouped within different domains of inquiry: *Demographic Information*, *Trial Experience*, *Acceptability and Knowledge of the Study Drug*, and *Sexual and Contraceptive Practices*. Within each domain, questions were given code names. These code names included a prefix for the domain and an identifier for the question topic. For example, within the *Trial Experience* section, participants were asked, "How do you feel about the HIV counseling sessions you received at the study clinic?" The code developed for this question was *TE_Counseling*. In each interview transcript, the code was applied to the section of text that included both the interviewer's question and the participant's response. Each question and respective response was coded this way. Once all of the data were structurally coded, the analysis team could easily sort through the data by question or domain to contextualize the data included in specific analyses (e.g., an analysis on trial experiences). Data from related questions could be easily consolidated and extracted from the full data set for further analysis.

Structural coding, as illustrated above, acts as a labeling and indexing device, allowing researchers to quickly access data likely to be relevant to a particular analysis from a larger data set. This makes analysis of copious data more efficient. Structural codes are also helpful for pulling together related data for development of data-driven thematic codes, but because they are developed without consulting the data, these codes are neither data-driven nor thematic in nature. Rather, as summarized by MacQueen et al. in chapter 6 of this volume, structural coding usually results in the identification of large pieces of text on broad topics that form the basis for an in-depth analysis within or across topics.

Frequencies

As noted in the description of content analysis earlier in this chapter, counting the frequency of a word or a phrase in a given data set is often a good starting point for identifying repeated ideas within a large body of text (Ryan and Bernard 2000:776), and gives an idea of the prevalence of thematic responses across participants. Simple keyword searches or word counts within a data set can allow a quick comparison of the words used by different subpopulations within an analysis (e.g., Ryan and Weisner 1996), or can be useful in developing a thematic codebook. So depending on what is being counted, frequencies can be part of either content or thematic analyses. The example provided below illustrates the advantage of simple keyword quantification as a potentially effective way to make decisions about what to include in or how to structure thematic codebooks.

In a study conducted in three West African countries, Namey and colleagues collected information on the experiences of people living with HIV/AIDS using semistructured interviews. One of the objectives of the research was to find out how people with HIV cope with the disease in resource-poor settings. The analysis was exploratory and focused on a short series of structurally coded questions in which respondents had been asked about their coping strategies. The analysis team compiled a list of the coping strategies identified by respondents as they read through the relevant structurally coded segments from forty transcripts from the three sites. In considering how to best code the strategies thematically, the team was faced with the following question: "Should a codebook be developed for each of the sites, or should the codebook be shared across all three sites?"

The team decided to use the MS Word search function in a qualitative data analysis (QDA) software package to answer the question. With researcher-defined parameters for the search, the program allowed them to identify the number of times each of the keywords associated with coping strategies appeared in the data sets from each of the sites. Examples of keywords used included *family*, *faith*, *doctor*, and *support group*. The search was limited to the segments of text structurally coded as part of the domain *Coping with HIV*. Within these segments, the frequency of the predetermined keywords was compared across the three sites. On the basis of this comparison, the researchers determined that the three sites were sufficiently similar to be analyzed using a single, shared codebook.

As the example above shows, word counts and keyword searches often require that researchers know what words to search for. Prior knowledge of the data is often essential, and accuracy of language is a concern. The data from the example above were originally collected in French, Pidgin English, and several local African languages, before being translated into English for analysis. The use of one specific word or another as a marker for an idea becomes less certain here, particularly without context to assist in checking meaning.

In the above example, the researchers focused the word count on the small section of text that was already known to relate to coping with HIV. This decision was made to avoid tallying uses of the common terms *family* and *doctor* that may have appeared in other contexts. However, the approach excluded any discussion of coping strategies that may have arisen spontaneously in other parts of the interview.

Another limitation to the use of content analysis is that keywords are acknowledged only as part of a simple word count, without taking the words' context into account. In the above example, both negative and positive mentions of *family* or *doctor* were counted equally, and without regard for whether the word actually described a coping strategy. This distinction would need to be made subsequently in a thematic codebook, after context from the transcripts was considered. Alternatively, "keyword in context" reports may be used to create a concordance that provides a certain number of words before and after the keyword, thus providing the context with which to interpret keyword frequencies (Ryan and Bernard 2000). The QDA Miner with WordStat and MaxQDA with MaxDictio software packages offer these and other content analysis capabilities.

Assessing the frequency of themes (i.e., code applications), rather than words, helps to incorporate context into the analysis, and is another helpful and fairly simple analytical technique. Different QDA software packages have different capabilities for creating code frequency reports, which list the number of times each of the codes from a selected thematic codebook is applied within a given data set. Some programs allow analysts to count the total number of files with the code applied, or the total number of times the code was applied across all files. These frequencies can also be tabulated without using software, though this can be quite time consuming.

Generally, we suggest determining frequencies on the basis of the number of individual participants who mention a particular theme, rather than the total number of times a theme appears in the text. We base this on the assumption that the number of individuals independently expressing the same idea is a better indicator of overall thematic importance than the absolute number of times a theme is expressed and coded. One talkative participant, for example, could express the same idea in many of her or his responses and thus significantly increase the overall frequency of a code application. However, the choice of frequency report ultimately depends on the analysis objective. In some cases the number of times a code was applied within each file—across different domains, for example—may be the best indication of the breadth or pervasiveness of a coded theme.

If coding has been done systematically using clearly defined codes, a code frequency report can help identify which themes, ideas, or domains were common and which rarely occurred. In this way, the number of times a code is applied can be used as an indication of the salience of a theme or an idea across files, domains, or questions, depending on the analysis objective. Coding frequencies, like word counts, can also be compared between different sources or subpopulations within the data to explore similarities and differences among these. When using code frequencies, it is important to keep in mind that the codes represent interpretive summaries of data, rather than the primary data themselves. This is another reason to monitor the reliability of the code application process.

Code frequency reports can also be helpful for managing data and for revising codebooks. For example, a code that tends to appear numerous times within a data set may need to be refined and, if any subthemes or subconcepts are identified, the analyst may need to create subcodes that describe these in more detail. Conversely, a code that rarely occurs may need to be redefined or eliminated from the codebook altogether. The "standard of care" example described below illustrates the use of code frequency reports.

MacQueen et al. (in press) used code frequency reports to frame their analysis of data about locally acceptable standards of care related to HIV prevention clinical trials. Key informants, health care providers, and potential clinical trial participants at ten sites in seven countries were identified to participate in semistructured interviews or focus groups. These respondents were presented with the description of a hypothetical clinical trial, and were then asked to evaluate three scenarios about ways to care for people who might become HIV-positive during the course of a clinical study. After each scenario, the participant was asked whether she or he considered the scenario fair treatment. In total, 130 individual interviews and twenty focus group discussions were conducted and transcribed. Predetermined codes of *fair*, *unfair*, and *mixed* (elements of fairness and unfairness together) were then applied to participants' responses to each of the three hypothetical care scenarios in the transcripts.

To determine the number of respondents who felt that a given scenario was fair, unfair, or of mixed fairness, the analysis team generated code frequency reports for each of the three codes and for each scenario, using AnSWR software (CDC 2004). In addition, more focused code frequency reports were generated to find out whether there were differences in response to the three scenarios across the different categories of respondents. In other words, key informants were compared to health workers, who were in turn compared to potential trial participants. In this way, the research team used the code frequency reports to identify the perceived level of fairness for each of the scenarios according to the type of respondent. In subsequent analyses, the team examined participants' specific reasons for considering the scenarios fair or unfair.

Co-occurrence

Code co-occurrence is defined as the application of two or more codes to a discrete segment of text from a unique respondent (Guest and McLellan 2003). For instance, a paragraph of text about strategies for coping with HIV extracted from one respondent's interview transcript may contain references to "family," "faith," and "doctors." Each of these references is a separate idea that requires a separate code. The paragraph would be coded with all three codes, which are then said to co-occur in this text segment.

Code co-occurrence reports often provide helpful information in understanding how thematic domains, concepts, or ideas are distributed within a data set, beyond simple frequencies. Patterns in data sets can be identified by examining co-occurrences such as correlation between code items, themes, respondents, or events (LeCompte and Schensul 1999). Some QDA software packages frame these types of reports in Boolean terms, allowing the analyst to search for the intersection of Code A AND Code B, or Code C NOT IN Code D. Again, which report is most helpful will depend on the analysis objective.

As with code frequencies, it may be instructive to look at the total number of times any two codes co-occur within individual data files as well as across the entire data set. For example, a unique pair of codes may co-occur hundred times within a data set of fifty files, but all hundred co-occurrences appear within a small number of individual data files only. Alternatively, a pair of codes might co-occur only twenty-five times, but be distributed across twenty-five different files within the data set. For this reason, it is often important to look at these two reports, code frequency and code co-occurrence, together, before giving weight to one co-occurrence or another.

In the "standard of care" example described above, MacQueen and Namey used code co-occurrence reports to help address why some categories of respondents considered certain hypothetical scenarios for treatment to be fair and others unfair. They created thematic codes describing different benefits or limitations of treatment (e.g., *drug access*, *counseling*, *long queues at the medical center*, or *poor service*), and looked at all of the co-occurrences of these codes with the fairness codes mentioned earlier (*fair*, *unfair*, *mixed*). The resulting co-occurrence report, when subdivided by scenario, indicated respondents' perceptions of benefits or limitations of research-related care in relation to their assessments of fairness. Again, the researchers used this information as a guide for reviewing and reorganizing the text data for interpretation.

Graph-theoretic Data Reduction Techniques

While traditional thematic or structured coding can be a first step in ordering large data sets, the richness of the various codes applied to the data, coupled with the possibility of having multiple salient themes, requires additional consideration. In particular, as summarized by Miles and Huberman (1994:69), analysts are often faced with the following questions: "How do all the codes and themes relate to each other? What is the big picture, and how does it relate to each theme or code? Where does one begin to tell the story?"

To answer some of these questions, graph-theoretic techniques, also referred to as semantic network analyses, may be used to identify complex semantic relationships in bodies of text (Barnett and Danowski 1992, Danowski 1993, Osgood 1959). Using co-occurrence matrices as input, these techniques graphically depict the relationship between semantic items, such as words or codes. The development of various software programs for the statistical analysis of qualitative data has allowed matrices to be analyzed with increasing sophistication (Borgatti 1996b, Doerfel and Barnett 1996, Schnegg and Bernard 1996, Weitzman and Miles 1995). The most common graph-theoretic techniques, discussed in more detail below, include multidimensional scaling and hierarchical cluster analysis.

The techniques described in the previous section are useful for reducing and/or guiding analysis of large amounts of text, although these approaches result in a lot of detail and require piecing together interpretations from multiple, distinct reports. Cluster analysis and multidimensional scaling (MDS) techniques provide a broader, more holistic perspective. These methods are designed to identify the "structures" of categories that fit a collection of observations. Rather than starting with a priori or emerging categorical structures and filling them with data, these techniques allow the analyst to identify the "natural groupings" within the data set. As such, these approaches constitute an important technique that can help generate hypotheses about relationships between concepts or ideas or confirm initial ideas about how the data fit together (Johnson and Wichern 2000).

Hierarchical Clustering Techniques

Hierarchical cluster analysis is an agglomerative methodology that identifies clusters of observations in a data set (Aldenderfer and Blashfield 1984). As described by Anderberg (1973), at its most elementary level, cluster analysis "sorts data into groups, such that the degree of natural association is high among members of the same group and low between members of different groups." In other words, a cluster analysis is a statistical method for grouping "like" things together. Before performing a cluster analysis, the analyst first needs to display qualitative data using similarity matrices. A similarity matrix is a table of scores that express numerical distances—or likeness—between data points. These matrices can be binary (using zeros and ones to indicate simply whether two data points are similar or not), or they can be valued (using a range of values to indicate the degree or strength of similarity). These matrices can be composed of data from various sources, including direct word counts, pile sorts, or code frequencies, depending on which item of analysis best addresses the analysis objectives. Table 7.1 below provides examples of two code-by-code matrices, the first binary, and the second a valued matrix, based on the aggregate binary matrices of several individual respondents.

Using these similarity matrices as input, cluster analysis allows a researcher to see patterns in large data sets. Moreover, a refined context can be created by careful selection of the words or codes to be included in a similarity matrix. If the words or codes are limited to those that appear

Table 7.1. Examples of Individual and Aggregated Similarity Matrices

A. Code 3 Code Similarity Matrix, Individual Case *(Binary Matrix)*

	Code A	*Code B*	*Code C*	*Code D*	*Code E*
Code A	—	0	1	1	0
Code B	0	—	1	0	0
Code C	1	1	—	0	1
Code D	1	0	0	—	0
Code E	0	0	1	0	—

In this matrix, we have used QDA software to determine whether (yes = 1, no = 0) specific codes co-occurred in the text file associated with one particular participant. Codes A and B, for example, did not co-occur, while Codes A and C did.

B. Code 3 Code Similarity Matrix, Aggregated Cases *(Value Matrix)*

	Code A	*Code B*	*Code C*	*Code D*	*Code E*
Code A	—	2	15	12	1
Code B	2	—	4	6	0
Code C	15	4	—	8	11
Code D	12	6	8	—	0
Code E	1	0	11	0	—

This matrix displays the sum of several individual matrices for the same codes assessed in Matrix A. Rather than determining simply whether the selected codes co-occurred, this matrix tells us the number of individual participant files in which two codes co-occurred. Codes C and E, for example, were found to co-occur in eleven participant files.

within a particular domain or structural-level code segment, a bounding context can be used in interpretations of the resulting clusters. The example below illustrates the use of hierarchical cluster analysis in an integrated analysis of a multisite qualitative data set.

Guest et al. (2005) used hierarchical cluster analysis to examine the risk reduction counseling experience of men participating in an HIV vaccine trial. Seven respondents were selected from each of five US sites, for a sample of thrity-five men in total. Responses to two questions about risk reduction counseling experience were coded with forty-six thematic codes using AnSWR QDA software. The researchers created a binary matrix in which the rows represented each of the thirty-five respondents and the columns represented each of the forty-six codes developed. Rather than determining the number of times the code was applied in the interview, the matrix was designed to identify only the presence of the code in each respondent's transcript, thus obviating the potential numerical bias that would occur if, for example, a code was repeatedly applied to a single respondent's interview. The initial matrix was produced in AnSWR, and then the resulting file was imported into ANTHROPAC (Borgatti 1996a) to conduct a hierarchical cluster analysis. For a detailed account of the procedures used, refer to Guest and McLellan (2003).

Figure 7.1 depicts the resulting dendrogram of a cluster analysis of men's risk reduction counseling experiences within the context of HIV

Figure 7.1. Hierarchical cluster analysis of thematic data.

vaccine trial participation. The figure illustrates broad clusters that break down into several smaller clusters. In this analysis, the unit of observation is the code, and the unit of analysis is the respondent. This means that all numbers, such as levels in the cluster, salience, and frequency counts, are references to the number of unique respondents. Within a cluster, there are codes that occur at a low rate or level and others that occur at higher levels. Codes grouped at higher levels suggest co-occurrence as well as high frequency within the text. Put another way, codes that occur at a high rate signify themes that lots of men talked about; codes in the same cluster signify different themes that tended to come up together in an interview. The dendrogram thus provides a starting point from which to develop a more comprehensive interpretation of the data.

As shown in figure 7.1, the coded data split into three main clusters. Two of these clusters are further divided into subclusters. To determine the cutoff level for defining a cluster, the research team compared the cluster tree diagram to code frequency and salience reports derived from the coded data in AnSWR. At level 4 (see bottom left of figure 7.1), Guest and McLellan found that more than 20 percent of respondents provided information associated with a code contained within a cluster. The tree diagram displays notable breaks in clusters for codes with a salience of level 4 or less, supporting the chosen benchmark.

From this predetermined cutoff point, Guest and McLellan filled in the details to build a "narrative" that was grounded in the data. They added relevant quotes from text, compared findings to the literature, and confirmed their interpretations with saliency and frequency tables. For example, the cluster on the right in figure 7.1 included several codes related to the level of rapport men reported as part of their risk reduction counseling experience. The codes comprising the center subcluster—No Hang Ups and Comfort Good codes—indicate a relationship between men's ability to talk freely about sexual behavior and comfort with the risk reduction counseling. The coded text corroborated and added context to this clustering, by providing the evidence from interview transcripts that men were accustomed to talking about their sexual behaviors with others and/or being part of a research project.

Similarly, in the cluster analysis, risk assessment questions were frequently connected to an increased self-awareness of risk behavior (subcluster on the left: Assessment of Risks and Aware Know codes), while in the text data, explicit vocalization of risk behavior was reported to be an eye-opening experience. The subcluster on the right includes the related elements of a positive trial experience and having a good counselor (Positive Experience and Good Counselor codes), which is supported by text describing the comfort and ease engendered by good rapport with a nonjudgmental and sincere counselor. In this way, the researchers moved between the cluster analysis and the text data to summarize their findings, demonstrating the creation of a narrative structured by the clustered configuration of the data and interwoven with quotes or details from the qualitative data.

Cluster analysis can also be used for text-based content analysis. Newly developed software, such as IN-SPIRE (Pacific Northwest National Laboratory 2004), can handle extremely large data sets (an estimated million words per hour), and using sophisticated clustering techniques can produce three-dimensional outputs of data configurations.

Multidimensional Scaling

Multidimensional scaling (MDS) is another useful graph-theoretic technique that is based on similarity matrices. Schiffman et al. (1981) describe MDS as "a powerful mathematical procedure which can systematize data by representing the similarities of objects spatially as in a map." For a set of observed similarities (or distances or proximities) between *N* pairs of items, MDS techniques allow the researcher to find a representation of the items in the fewest possible dimensions in space, such that the distances between the points match as well as possible the original distances or proximities (Johnson and Wichern 2002).

A classic heuristic for explaining MDS can clarify this explanation. Think about creating a map using only a set of distances between various geographical locations. Given only the distances between all pairings of ten cities from around the borders of the United States, any MDS program (in SPSS, Systat, or SAS, for example) would draw a reasonable outline map of the country. This capacity of MDS to create visual, spatial representations of relationships among data points is particularly useful for more abstract data, like themes or ideas, which have no material or geographic representation. Ideas that are judged to be closer to each other according to the distance between them (similarity) are represented as points that are close to each other, and locations that are judged to be far from each other are represented as points that are distant, just as on a map (Schiffman et al. 1981). This "map" then serves as a guide for further analysis and interpretation of the text data from which it was derived.

Once an MDS plot is created, a measure of "stress" (which ranges from 0 to 1) is used to evaluate the extent to which the original input distances match the resulting diagram. "Stress" therefore refers to the amount of distortion between the actual numerical values in a proximity matrix and the representation of those values in the MDS plot (Clark et al. 1998). Zero stress indicates a perfect fit of the data, so the closer the "stress" is to zero, the better the fit of the MDS diagram. A traditional rule of thumb is that stress below 0.1 indicates a good fit between the actual proximities and their representation in the plot (Kruskal 1964). Note, though, that guidelines in this regard are evolving (see Sturrock and Rocha 2000).

As with hierarchical clustering, data designed to be analyzed using MDS can be entered into ANTHROPAC software (Borgatti 1996a) or other statistical software packages such as SPSS. These data can originate from words or thematic codes created in the process of qualitative analysis of text data, or from less traditional qualitative methods like pile sorting or triadic comparisons (e.g., Weller and Romney 1988). Data are first converted to proximity matrices, in a procedure similar to that outlined in the example from Guest and McLellan above, and are then analyzed using MDS. To keep things relatively simple, the following example is based on a rather small sample of pile sort data, collected in one location; however, the same methods of data preparation and analysis are applicable to, and even more useful for, larger sets of data collected in multiple sites.

Thairu (2005) asked a purposive sample of women from the Pemba Island of Tanzania to sort thirty-four cards, each containing a short description of a common caregiving practice for newborns in the community. To ensure that responses across participants were comparable, she restricted the number of piles that respondents could make to four. If an informant put any two items in the same pile, the items were considered to be similar; conversely, if the informant put any two items in different piles, the items were judged to be different (Weller and Romney 1988). In preparing the data obtained for analysis, each informant's similarity judgments were represented as an item-by-item matrix in which each cell contained either 1 for items that appeared in the same pile or 0 for items that appeared in different piles (Borgatti 1996b). One matrix represented each informant.

Thairu calculated the aggregate similarity for any two items, Item A and Item B, by adding the values in Cell AB across all *N* informants and dividing by *N*. The resulting number was the percentage of informants in the sample who placed Item A and Item B in the same pile. The aggregate similarity matrix served as the input for multidimensional scaling. Using ANTHROPAC (Borgatti 1996a), she compared similarities among the items and found a set of points in space such that the distances among these points corresponded as closely as possible to the input proximities.

Table 7.2 shows the level of stress (or fit) of the model created with the data, using an increasing number of dimensions. As shown in the table, increasing the number of dimensions reduced the level of stress. However,

Table 7.2. Stresses Obtained from the Multidimensional Scaling Algorithm

Dimensions	*Stress Obtained*
1	0.425
2	0.243
3	0.143
4	0.098
5	0.068

increasing the number of dimensions above two makes it difficult to display on paper, and even more difficult to comprehend. As Borgatti (1996b) argues, beyond four or more dimensions, MDS does not make complex data accessible to the human mind. For this reason, while recognizing that adding more dimensions would provide a more accurate geometric representation, Thairu chose to stop at three dimensions with a stress of 0.143. The MDS plot in figure 7.2 shows the spatial clustering of items from the pile sort activity, in two dimensions.

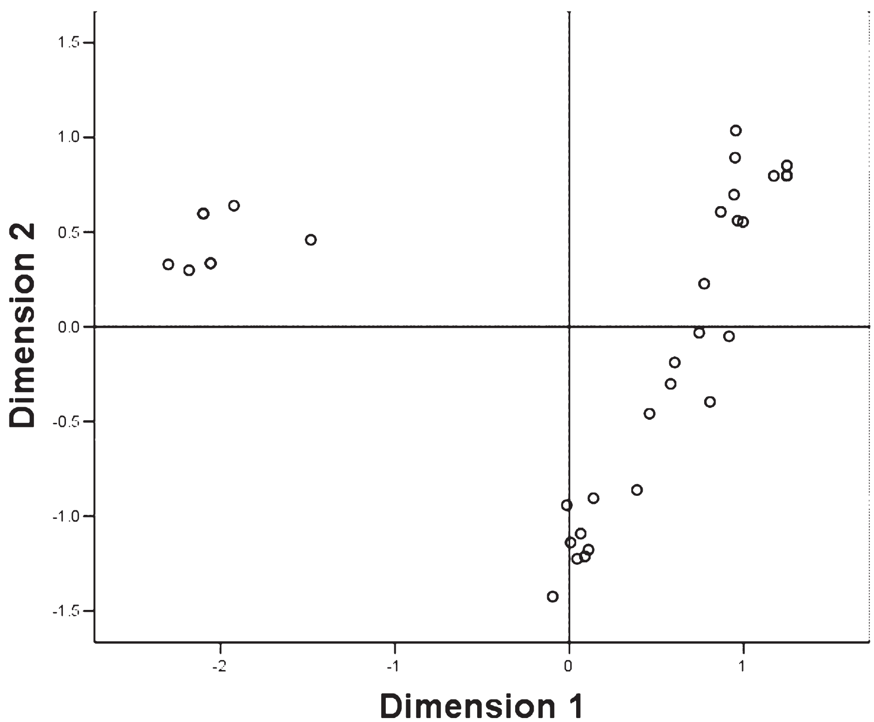

Figure 7.2. MDS plot of judged similarities.

In the analysis, one grouping of behaviors was observed to be located at the upper left extremity of the two dimensions, while other behaviors formed an ellipse spanning the right extremity of the two dimensions. The clusters were loosely grouped and interpreted as shown below in figure 7.3, with specific items detailed in table 7.3.

According to Thairu, in the diagram, the first dimension allowed a separation between the infant feeding cluster and the other cluster of behaviors. The second dimension represented a continuum of "time" indicating the behaviors respondents undertook following birth and as the infant grew older.

Thairu noted that there was considerable ambiguity and looseness in the groupings she found with respect to the rationale for some of the behaviors. For example, for many of the women interviewed, the practices of *calling the baby to prayer* and *applying soot on the baby's feet, palms, and face* were classified together because they were perceived to be "traditional" behaviors. In discussing their grouping decisions for these behaviors, women commented on their traditionality. On the other hand, some of the women

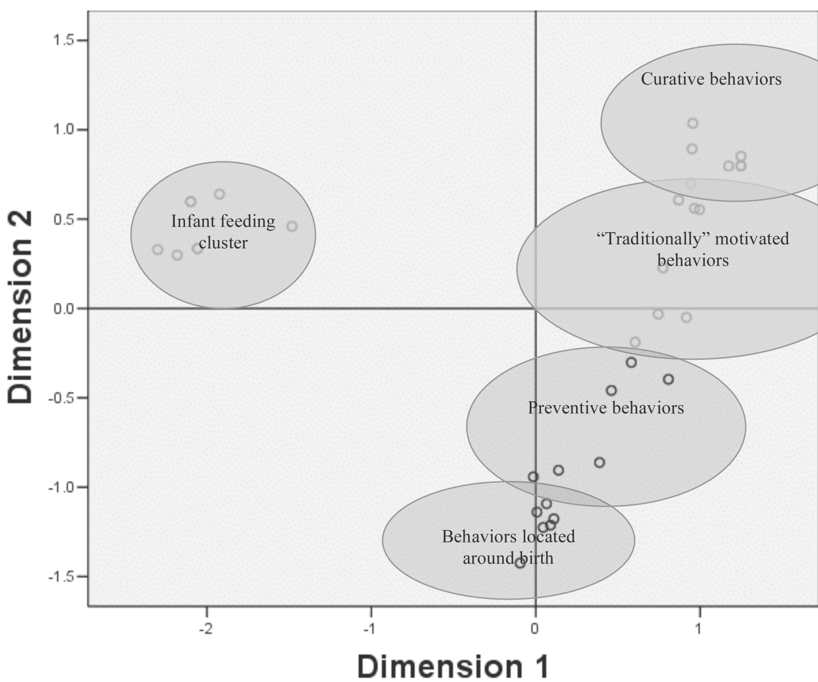

Figure 7.3. Clusters identified through MDS scaling of the peri-urban women's sorting of the thirty-four newborn care behaviors.

Table 7.3. Specific Behaviors Contained in Each of the Clusters

Cluster Label	*Specific Items Grouped within the Cluster*	
Feeding behaviors	Baby given liquids using a spoon Baby given liquids using a bottle Baby given biscuits	Baby given powder milk Baby given water Baby given cow's milk Baby given porridge Baby breast-fed
Preventive behaviors	Baby wrapped Baby wrapped with a khanga Baby washed with water Baby washed with soap Oil applied on baby's skin Olive oil applied on baby's skin Oil applied on fontanelle Baby massaged with warm hands	Baby massaged with coconut oil Oil used to massage baby does not have a strong smell Soap used to wash baby does not have a strong smell Baby washed with Swahili medicine
Curative behaviors	Part of baby's hair shaved and medicine applied on shaved part Baby not washed	Baby made to inhale smoke from burned Swahili medicine
"Traditionally" motivated behaviors	Baby called to prayer Soot applied on baby's feet, palms, and face Kohl applied on baby's forehead or eyes	Oil mixed with Swahili Medicine applied on baby's skin Baby passed the doorway seven days after birth
Birth-related behaviors	Giving birth at home Giving birth in the hospital Baby vaccinated	Razor used to cut the cord boiled Cord tied using a string or thread Placenta buried

chose to group the same behaviors as preventive care practices that are undertaken to avert unfavorable outcomes for the infant.

Not surprisingly, most women in Thairu's sample could not provide clear labels or concepts for their sorting, so she had to create labels that reflected her understanding of the emic picture. This inability to articulate cultural knowledge illustrates the utility of cognitive mapping exercises, and also highlights the care that must be taken in interpreting them. Prior familiarity with the data, context, and research problem is a prerequisite to accurate interpretation, since this knowledge enables the researcher to distinguish between "good" and "bad" groupings when confronted with them.

Comparing Qualitative Data

Many large qualitative research projects are designed to explore similarities and differences between two or more sites, groups of people, or other dimensions of interest. Such an objective often requires direct comparisons of data, yet comparing qualitative data in a meaningful way can be challenging, given the unstructured nature of responses. Certainly, a semistructured instrument—in which similar or identical questions are asked of all respondents, preferably in the same order—facilitates comparison, but certain issues remain. Even if, for example, a semistructured instrument is used, is it meaningful to say that theme *x* was identified in text from 75 percent of the participants in one group and only 25 percent in a comparison group? Or, what if a particular theme (or specific word, in the case of classic content analysis) is present in 20 percent of the transcripts from one site and in none of the transcripts from the other site? At what point do we say there is a meaningful difference between the two groups?

One possibility is to construct cross-tabs and carry out nonparametric procedures on frequencies, such as chi-square or Fisher's Exact Test (see chapter 11, this volume). Because these are nonparametric approaches, sampling requirements are not as rigid as for those associated with other statistical tests, and are suitable to the types of studies that employ qualitative methods. However, their use comes with some caveats, as some issues arise that are unique to comparing open-ended responses.

The response possibilities for an open-ended question are infinite, as opposed to finite categories in a fixed-response question. And, in the case of fixed-response questions, a particular answer is mutually exclusive of other response categories. The same cannot be said for responses to open-ended questions. We therefore have to assume that the same conditions are acting upon the responses across all of the groups being statistically compared. This is often not the case in team research, where interviews are sometimes conducted by different team members. Despite the best training efforts and the use of relatively structured interview guides, different interview styles and probing techniques will emerge between team members, thereby producing different responses and biasing the pattern of responses.

That said, if an analysis plan calls for the use of nonparametric statistics and multiple interviewers are to be employed, we recommend that the interviewees are distributed randomly (or at least evenly with respect to interview group) among interviewers. If random or even allocation of interviews is not feasible, all interviewers should at least have a chance to interview a few participants from each group to be compared; the subsequent chi-square analysis can then be stratified to control for interviewer effect.

The presence of a theme in one group of respondents and its absence in a different group of respondents is likely meaningful, and should be reported and interpreted accordingly. Graphical displays of frequencies relative to each other can also help in interpreting comparative data. Using a technique derived from MacQueen et al. (2001), Guest et al. (2005) created a simple line graph showing the relative percentage of participants expressing key themes for respondents of two different age groups (figure 7.4). While a greater percentage of the younger group expressed the key themes, for the two age groups the relative frequency among themes was virtually identical. The graphs indicate that the themes held the same relative importance for each group, but that the younger group was more vocal about the subject matter overall.

Co-occurrence of data elements can also be compared in graphical form. Separate cluster analyses or MDS outputs can be created for particular groups of interest and compared and interpreted qualitatively.[1] In the Pemba island example, Thairu subsequently compared the results of the newborn care pile sorting by peri-urban women to the results of a sample of rural women. She found striking similarities in both samples: the infant feeding grouping remained separate from other clusters of behaviors, and the other clusters were not distinctly separated. However, for the rural women, the feeding cluster was located at the bottom left of the two-dimensional graphical representation in contrast to the top left for the peri-urban women. This indicated subtle differences between the two samples' conceptualizations of infant feeding, which were confirmed in the thematic analysis of interview transcripts.

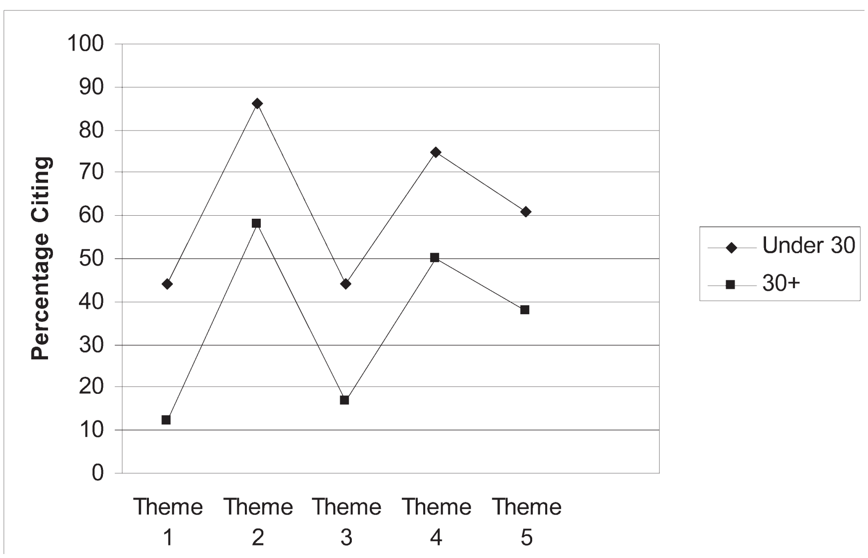

Figure 7.4. Age differences and thematic frequency.

Peri-urban

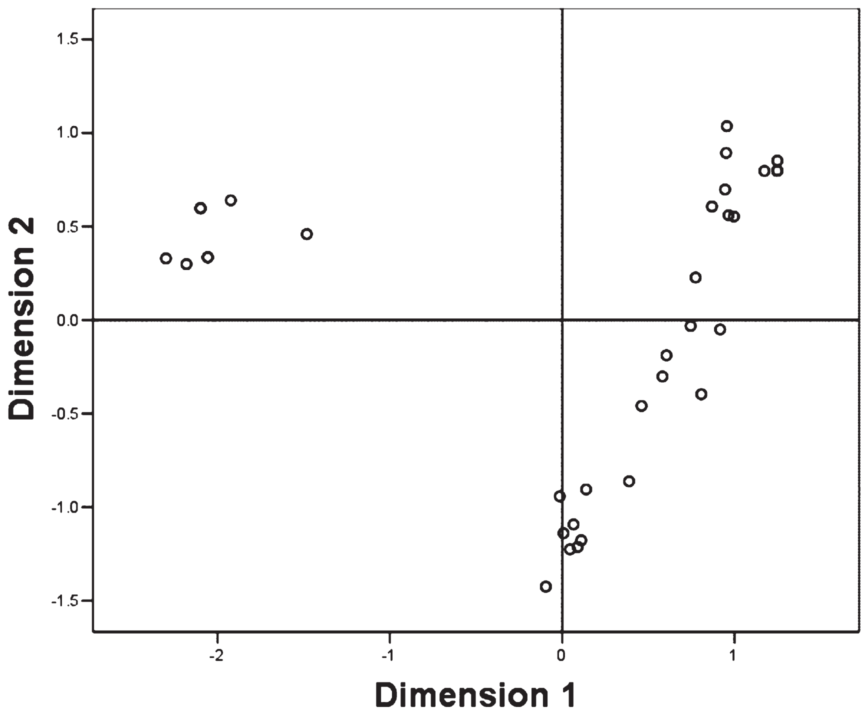

Figure 7.5a&b. Comparison of clusters identified through MDS scaling of rural and peri-urban women's sorting of thirty-four newborn care behaviors.

Qualitative Data Analysis (QDA) Software

For many (but not all) of the analytical techniques we have described above, QDA software technology is essential. Choosing an appropriate software package for a particular analysis task depends on the researchers' analysis objectives, as well as the size and complexity of both the data set and the analysis team (see MacQueen 2005). Lewins and Silver (in press) present a helpful list of questions to ask before choosing a qualitative data analysis software package, and provide reviews of several commonly used software systems, including ATLAS.ti 5, HyperRESEARCH 2.6, MAXqda2 (MAXdictio & MAXmaps), N6, NVivo2, NVivo7, QDA Miner, Qualrus, and Transana. For up-to-date reviews of the capabilities and limitations of different software packages, we suggest that the reader visit the CAQDAS (pronounced "cactus") Networking Project website, at caqdas.soc.surrey.ac.uk.

Conclusions

There is no single "right" way to approach analysis of a large qualitative data set, and often an assortment of complementary approaches, building one upon another and triangulating findings, is preferable. As discussed in this chapter, researchers may use content analysis of words or phrases as an initial method to filter the data and reduce the data set, or they can undertake structural coding of an entire data set before proceeding with further thematic analyses. Once data have been coded for themes, we have proposed checking codes for co-occurrences and code frequencies, or using the codes in similarity matrices to produce graphical data displays. A combination of these techniques may be particularly useful for researchers working on team-based and/or multisite projects, to help simplify large amounts of textual data. Simplification allows researchers to focus their attention on the rich, descriptive, and expressive details of qualitative data, without getting lost among those details.

As sociobehavioral research designs become larger and more complex, qualitative researchers have a greater opportunity to test and expand the range of methods in their repertoire to make the best use of resources and data generated by multisite, team-based research. We hope that the examples presented in this chapter serve as a starting point for wider exploration of innovations in qualitative data analysis.

Note

1. For additional ideas for visually displaying data, we recommend looking through Edward Tufte's colorful trilogy: *Envisioning Information* (1990), *The Visual Display of Quantitative Information* (1992), and *Visual Explanations: Images and Quantities, Evidence and Narrative* (1997). All three books provide numerous and unique examples of how data can be brought to life and made intuitive through graphical representation.

References

Aldenderfer, M. S., and R. K. Blashfield
1984 *Cluster Analysis.* Thousand Oaks, Calif.: Sage.

Anderberg, Michael R.
1973 *Cluster Analysis for Applications.* New York: Academic.

Barnett, G., and J. Danowski
1992 The structure of communication: A network analysis of the International Communication Association. *Human Communication Resources* 19:164–285.

Bernard, H. R., and G. Ryan
1998 Text analysis: Qualitative and quantitative methods. In H. R. Bernard, ed. *Handbook of Methods in Cultural Anthropology*. Pp. 595–645. Walnut Creek, Calif.: Altamira.

Borgatti, S.
1996a ANTHROPAC 4.0. Natick, Mass.: Analytic Technologies.
1996b 4.0 Methods Guide. Natick, Mass.: Analytic Technologies.

Centers for Disease Control and Prevention
2004 *AnSWR: Analysis Software for Word-based Records, Version 6.4.* Altanta, Ga.: Centers for Disease Control and Prevention.

Clark, L., C. Vojir, N. O. Hester, R. Foster, and K. L. Miller
1998 MDS and QAP: How do children rate painful experiences? In V. C. de Munck and E. J. Sobo, eds. *Using Methods in the Field: A Practical Introduction and Casebook*. Walnut Creek, Calif.: AltaMira Press.

Danowski, J.
1993 Network analysis of message content. In D. Richards and G. Barnett, eds. *Progress in Communication Science, XII*. Pp. 197–221. Norwood, N.J.: Ablex.

Dey, I.
1993 *Qualitative Data Analysis: A User-friendly Guide for Social Scientists.* New York: Routledge.

Doerfel, M., and G. Barnett
1996 The use of Catpac for text analysis. *Cultural Anthropology Methods Journal* 8(2):4–7.

Glaser, B., and A. Strauss
1967 *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine.

Guest, G., and E. McLellan
2003 Distinguishing the trees from the forest: Applying cluster analysis to thematic qualitative data. *Field Methods* 15(2):186–201.

Guest, G., E. McLellan-Lemal, D. M. Matia, R. Pickard, J. Fuchs, D. McKirnan, and J. L. Neidig
2005 HIV vaccine efficacy trial participation: Men who have sex with men's experiences of risk reduction counseling and perceptions of risk behavior change. *AIDS Care* 17(1):46–57.

Johnson, R. A., and D. W. Wichern
2002 *Applied Multivariate Statistical Analysis.* Upper Saddle River, N.J.: Prentice Hall.

Kearney, M., S. Murphy, and M. Rosenbaum
1994 Mothering on crack cocaine: A grounded theory analysis. *Social Science and Medicine* 38:351–361.

Krippendorf, K.
1980 *Content Analysis: An Introduction to Its Methodology.* Beverly Hills, Calif.: Sage.

Kruskal, J. B.
1964 Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1-27.

LeCompte, M. D., and J. J. Schensul
1999 *Analyzing and Interpreting Ethnographic Data*. Walnut Creek, Calif.: Altamira.

Lewins, A., and C. Silver
2007 Choosing a CAQDAS package. In A. Lewins and C. Silver, eds. *Using Software in Qualitative Research: A Step-by-Step Guide*. London: Sage.

MacQueen, K. M.
2005 What to look for in software for qualitative data analysis. In Priscilla R. Ulin, Elizabeth T. Robinson, and Elizabeth Tolley, eds. *Qualitative Methods in Public Health: A Field Guide for Applied Research*. Pp. 172–174. San Francisco: Jossey-Bass.

MacQueen, K. M., E. McLellan, D. Metzger, S. Kegeles, R. Strauss, R. Scotti, L. Blanchard, and R. Trotter
2001 What is community? An evidence-based definition for participatory public health. *American Journal of Public Health* 91:1929–1937.

MacQueen, K. M., and B. Milstein
1999 A systems approach to qualitative data management and analysis. *Field Methods* 11: 27–39.

MacQueen, K. M., E. Namey, D. Chilongozi, S. Mtweve, M. Mlingo, N. Morar, C. Reid, A. Ristow, S. Sahay, and the HPTN 035 Standard of Care Assessment Team
In press Community perspectives on care options for HIV prevention trial participants. *AIDS Care*.

Miles, M. B., and M. A. Huberman
1994 *Qualitative Data Analysis: An Expanded Sourcebook* (2nd ed.). Thousand Oaks, Calif.: Sage.

Osgood, C.
1959 The representational model and relevant research methods. In I. de Sola Pool, ed. *Trends in Content Analysis*. Pp. 33–88. Urbana: University of Illinois Press.

Pacific Northwest National Laboratory
2004 *IN-SPIRE, 2.3*. Seattle, Wash.: Pacific Northwest National Laboratory.

Patton, M. Q.
1990 *Qualitative Evaluation and Research Methods* (2nd ed.). Newbury Park, Calif.: Sage.

Ryan, G., and R. Bernard
2000 Data management and analysis methods. In Norman Denzin and Yvonna Lincoln, eds. *Handbook of Qualitative Research.* Pp. 769–802. Thousand Oaks, Calif.: Sage.

Ryan, G., and T. Weisner
1996 Analyzing words in brief descriptions: Fathers and mothers describe their children. *Cultural Anthropology Methods Journal* 8:13–16.

Schiffman, S. S., L. M. Reynolds, and F. W. Young
1981 *Introduction to Multidimensional Scaling:Theory, Methods, and Applications.* New York: Academic.

Schnegg, M., and H. R. Bernard
1996 Words as actors: A method for doing semantic network analysis. *Cultural Anthropology Methods Journal* 8(2):7–10.

Sturrock, K., and J. Rocha
2000 A Multidimensional Scaling Stress Evaluation Table. *Field Methods* 12(1):49–60.

Thairu, L.
2005 Ethnographic case studies of infant feeding in the context of HIV/AIDS and newborn care in sub-Saharan Africa. Ph.D. dissertation, Division of Nutritional Sciences, Cornell University, Ithaca, N.Y.

Tufte, E.
1990 *Envisioning Information.* Cheshire, Conn.: Graphics Press.
1992 *The Visual Display of Quantitative Information.* Cheshire, Conn.: Graphics Press.
1997 *Visual Explanations: Images and Quantities, Evidence and Narrative.* Cheshire, Conn.: Graphics Press.

Weber, R. P.
1990 *Basic Content Analysis.* Newbury Park, Calif.: Sage.

Weitzman, E. A., and M. B. Miles
1995 *Computer Programs for Qualitative Data Analysis.* Thousand Oaks, Calif.: Sage.

Weller, S., and A. Romney
1988 *Systematic Data Collection.* Newbury Park, Calif.: Sage.

Wright, K. B.
1997 Shared ideology in Alcoholics Anonymous: A grounded theory approach. *Journal of Health Communication* 2(2):83–99.

III

QUALITY CONTROL AND ASSURANCE

A real understanding of Quality doesn't just serve the System, or even beat it or even escape it. A real understanding of Quality captures *the System, tames it, and puts it to work for one's own personal use, while leaving one completely free to fulfill his inner destiny.*

ROBERT M. PIRSIG, ZEN AND *THE ART OF MOTORCYCLE MAINTENANCE*