Distinguishing the Trees from the Forest: Applying Cluster Analysis to Thematic Qualitative Data

GREG GUEST
ELEANOR MCELellan
Centers for Disease Control and Prevention

Qualitative data analysis requires organizing and synthesizing often large quantities of text. In many cases, this analysis entails negotiating the interplay between raw data, semantic themes or codes, and the overarching conceptual framework. In this article, the authors use a case study, which examines HIV vaccine efficacy trial participants' discourse, to demonstrate how cluster analysis can be used to aid in the analysis of large qualitative data sets. After briefly reviewing the systematic approaches to qualitative analysis and describing the project background, the authors present an example of how a hierarchical cluster technique can be incorporated into a multistage thematic analysis.

Keywords: cluster analysis; thematic analysis; coding

The analysis and interpretation of textual data can be an unwieldy task for even the most experienced of analysts. The continued development and refinement of analytical techniques, along with advancements in qualitative data analysis (QDA) software, have greatly assisted the process. However, the ability to connect numerous details and simultaneously formulate a more comprehensive understanding of a qualitative data set remains a challenge.

One problem has to do with gaining perspective on an entire data set after coding is completed. Building a codebook and applying codes to a body of text require repeated exposure to the raw data, essentially grounding one in the data. Undoubtedly, such exposure increases a researcher’s familiarity with the text and its subtleties, but such intimate knowledge may also obfuscate the bigger picture, the proverbial inability to see the forest for the trees.

The complexity and the often fuzzy nature of qualitative data present additional challenges. In a robust and well-coded qualitative data set, numerous codes and salient themes invariably emerge. Such robustness, in turn, creates quandaries for the analyst in interpreting and presenting findings.

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? Faced with numerous salient themes, where does one begin to tell the story? These are quintessential questions that arise when one is performing and reporting qualitative analyses (Miles and Huberman 1994:69).

In this article, we present a case study from HIV vaccine preventive research to illustrate the usefulness of cluster analysis in framing data interpretation. In particular, we demonstrate how cluster analysis can aid in visualizing the overall relationship among a multitude of codes and themes and to corroborate interpretation based on other analytical strategies such as code frequencies and isolated pairs of co-occurrence. Details of coding and coder reliability are not addressed here, nor do we provide a review of the various software packages that support this technique. Instead, we describe a multi-stage thematic analysis process specifically focusing on the application of a cluster technique. We begin with a brief review of structured QDA methodology.

A positivist approach to qualitative data focuses on reducing text to codes and applying “quantitative methods to find patterns in the relations among the codes” (Bernard and Ryan 1998:596). During the past few decades, researchers have developed a host of methods for structuring and analyzing textual data (for a review of these, see Dey 1993; Bernard and Ryan 1998; LeCompte and Schensul 1999). At a basic level is content analysis, which centers on determining the frequency of particular words or phrases in the text. Word counts can be expanded to include associated attributes of keywords, such as synonyms, location in the text, and surrounding words or phrases (Dey 1993:59).

Another strategy is a thematic analysis, in which codes are applied or linked to the raw data. This approach can be data driven, as in grounded theory (Glaser and Strauss 1967; Kearney, Murphy, and Rosenbaum 1994; Wright 1997), or more theory driven, as exemplified in classic content analysis (Krippendorf 1980; Weber 1990). Such analyses are useful for comparing the relative frequency, and arguably the salience, of items of analysis but do not tell us much about the patterns among the items. As Richards and Richards (1995) pointed out,

Categorizing is never just an end in itself, and is rarely recommended merely as a means of tagging all data on a topic. Rather, its goals are often the discovery and ordering of ideas and themes; and the storing of growing understandings, the linking of ideas to data, cross-referencing, sorting and clarifying. Qualitative researchers are urged not merely to derive and use categories but to do so in particular ways. (P. 80)
Patterns can be identified by examining co-occurrences such as correlation between code items, themes, respondents, or events (LeCompte and Schensul 1999). Similarly, Boolean minimization techniques, by which items are arranged on the basis of a limited set of logical relationships (e.g., if A and not B, or if B and C), although not quantitative, can reveal types of patterns between themes and codes (Ragin 1995; Romme 1995). Saliency (i.e., the number of times a code appears in all combination of codes) or typicality (Ryan 1999) can also indicate the centrality of particular themes in a body of text. These analyses, although useful, do not reveal the more explicit hierarchical relationships among codes or themes.

Richards and Richards (1995) divided the various ways of integrating hierarchy into a data set into two methodological camps: data driven (bottom-up) and theory driven (top-down). Taxonomic trees are typically created at the conceptual level (i.e., the researcher manipulates themes hierarchically according to a set of research objectives or in response to the perceived logical relationships among codes). Most QDA software allows the user to manipulate relationships among codes and create taxonomies or scales in the digital environment.

These methods of analysis are indeed useful but may limit the ability to extract the larger picture, one that takes into account the interrelationship between the individual codes as they occur within the data. Code co-occurrence, for example, does little to reveal the hierarchical relationships of themes. Conversely, the hierarchical methods described earlier do not necessarily capture patterns between themes in the text but rather create taxonomies based on the researcher’s subjective interpretation of logically consistent conceptual relationships between themes and codes.

Graph-theoretic techniques address the issues of subjectivity and hierarchy in semantic analyses to a certain degree. Often referred to as semantic network analyses, graph-theoretic approaches were developed as early as 1959 to identify complex semantic relationships in bodies of text (Osgood 1959; Barnett and Danowski 1992; Danowski 1993). Using co-occurrence matrices as input, these techniques graphically depict the relationship between semantic items, such as words or codes. With the increased use of computers and QDA software (Weitzman and Miles 1995), matrices have been treated with increasing sophistication (Borgatti 1996; Doerfel and Barnett 1996; Schneeg and Bernard 1996). One example is multidimensional scaling (e.g., Jang and Barnett 1994). Although multidimensional scaling is grounded in the structure of the data, the relationship between items is often opaque and difficult to see clearly. The demarcation of groups of items can be problematic, in particular when lengthy item labels are involved. In addition, interpreting several conceptual dimensions at once can be overwhelming.

Cluster analysis provides a useful alternative as it presents data in clearly defined clusters in two-dimensional space, rendering a quick and easy visual tool for interpretation. The analytical process we describe, which applies cluster analysis to thematic codes, was implemented by MacQueen et al. (2001:1930) for large, multisite qualitative studies. We use the following case study, from HIV vaccine preventive research, to more explicitly demonstrate the application of the technique and its usefulness in the interpretation and presentation of qualitative data.

CASE STUDY

Background

The first large-scale HIV vaccine efficacy trial in North America was initiated in June 1998. The three-year, randomized, double-blind, placebo-controlled trial was sponsored by VaxGen, Inc., with support from the Centers for Disease Control and Prevention and the National Institute of Health. The trial, conducted in sixty-one sites in the United States, Canada, and the Netherlands, involved 5,417 HIV-uninfected individuals at risk for the sexual transmission of HIV. Ninety-four percent of these volunteers were men who have sex with men. Centers for Disease Control and Prevention staff and investigators from five domestic clinical sites conducted supplemental biomedical and behavioral research within the larger trial.

Part of this supplemental research involved a qualitative study. The rationale for undertaking a qualitative study was based on the likelihood that several additional, large, phase III efficacy trials would be required during HIV vaccine development. Hence, it was imperative that researchers more fully understand the personal motivations and the social context of HIV vaccine trial participation.

A key topic related to HIV vaccine preventive research is the effect, if any, that regularly scheduled risk-reduction counseling has on participants. Initial HIV vaccine feasibility studies indicated that in spite of a double-blind design, trial participants are likely to engage in riskier sexual behavior because of a false perception that they are being protected by the vaccine (MacQueen et al. 1994). We analyzed participants’ experiences related to risk-reduction counseling and their perceptions of their changes in risk behavior change during the vaccine trial.

Our sample consisted of thirty-five male trial participants, all of whom reported having sex with other men. The interview guide contained ten open-ended questions that specifically addressed risk-reduction counseling and
perception of behavior change. The responses were transcribed verbatim into computer files, coding was performed in AnSWR: Analysis Software for Word-Based Records (Strovan et al. 2002), and the interpretation of coded data was handled in both AnSWR and ANTHROPAC (Borgatti 1996).

Data Analysis and Interpretation

Using a standardized iterative process, we developed a structured codebook (MacQueen et al. 1998) and coded the text in AnSWR. To support examination of the final coded data, a structural code was created for each domain of inquiry to provide a context within which to analyze and interpret responses. Domains of inquiry are derived from the research objectives and are composed of one or more interview questions. Structural codes, which represent domains of inquiry (or subsets of domains), were established prior to analysis and were applied to segments of text that included both the interview question(s) within a given domain of inquiry and the accompanying response(s). For our analysis, each text segment was associated with a single structural code. A structured, open-ended interview approach made it possible to apply a mutually exclusive structural code to all thirty-five interviews. Theoretically, multiple structural codes can be applied to a text segment; however, data management and retrieval become increasingly complex. Moreover, it may introduce an analytical framework that is top heavy and a codebook that is cumbersome to use.

Although thematic coding can be completed without structural coding, the use of structural codes supports greater flexibility in refining analyses by context, in particular when dealing with interview scripts that cover a large number of conceptual domains. As the purpose of this article is to illustrate the application of cluster analysis to thematic codes, we present an analysis based on only one domain of inquiry, overall experience of risk-reduction counseling (RRC Experience structural code), which was composed of the two questions below. It should be noted that in some cases, a domain of inquiry containing multiple questions may produce responses that do not overlap and that therefore require more of a hierarchical structural coding approach with development of individual content-driven codes for each question. For the case study presented, respondents did not provide responses that differed for the two questions within the domain of RRC Experience.

Tell me about your experiences and thoughts about the visits that focused on your risk behavior.

Describe what it was like talking about your risk-taking behavior.

After structural coding had been completed, two persons independently reviewed the transcripts and identified key concepts associated with participants’ responses. A code was assigned to each of these data-driven concepts, text segments were coded for content, and intercoder agreement was assessed. To minimize data overload, the text was coded in batches of segments. We assessed intercoder agreement cumulatively, as well as for each individual code, using kappa scores (Carey, Morgan, and Oxtoby 1996). The analysis team discussed coding discrepancies, the codebook was revised accordingly, and recoding was performed when necessary. After several passes through the text and numerous coding iterations, we ended up with a total of forty-six content-derived codes associated with the RRC Experience structural code (see Table 1). Once the coding was completed, the structural codes were used to examine the data in context so that we could then describe and make sense of how participants had constructed their experiences and perceptions.

We used a multistage analysis process to interpret the thematic coding and used AnSWR-reporting functions to generate the following output:

- code frequencies—number of times a code was applied to a text segment and number of times that a code was associated with a unique respondent,
- code co-occurrences—isolated pairs of codes applied to text segment and associated with a unique respondent, and
- saliency—number of times that a code occurred within a combination of codes delineated by either text segments or respondents (note that, in theory, a code that repeatedly occurs in isolation may exhibit high frequency but have low to no saliency).

As informative as these reports were, it was difficult to visualize the larger pattern. Reliance on multiple individual reports makes it challenging to piece together the results. At this point in the process, it is difficult to see the larger schematic framework. Moreover, having reviewed the raw data so many times, we had gained an intimate knowledge of the text and its thematic content, but such knowledge can be an impediment. Thematic associations can be complex and numerous, and in the absence of a more objective method of data reduction, interpretation may reflect subjective salience or may result from selective memory effects. Often, what is needed is a method for gaining distance from the data. The cluster analysis provides a quick and effective means of data reduction that is both meaningful and easy to read.

The first step in preparing the coded data for cluster analysis was to generate a binary \((m \times n)\) matrix. We selected a binary matrix in which the respondents represent the rows in the matrix and the codes are the columns because
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptable risks</td>
<td>Perception or attitude that one can establish sex boundaries that permit one to be safe yet engage in behaviors that he enjoys</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>Statements that refer to information received in counseling as being ambiguous or contradictory to information received elsewhere</td>
</tr>
<tr>
<td>Assessment of risks</td>
<td>Talk about risk assessment (as opposed to RRC) and techniques used to figure out an individual’s risk-taking behavior</td>
</tr>
<tr>
<td>Aware know</td>
<td>Perceptions or attitudes that trial participation has increased or can increase awareness of one’s risk behaviors, including a self-risk inventory and recognition of acceptable risks</td>
</tr>
<tr>
<td>Behavior worse</td>
<td>Statements that behavior would have been worse if trial participation had not taken place, or that behavior was worse prior to trial participation</td>
</tr>
<tr>
<td>Boring</td>
<td>Perception or attitude that safe sex is boring and/or limits pleasure</td>
</tr>
<tr>
<td>Comfort good</td>
<td>Statements that participant was comfortable with RRC</td>
</tr>
<tr>
<td>Comfort low</td>
<td>Limited or negative comfort level with RRC</td>
</tr>
<tr>
<td>Comfort shift</td>
<td>Explanation that initial level of discomfort shifted or dissipated after becoming better acquainted with the trial counselor(s) and trial procedures</td>
</tr>
<tr>
<td>Condom access</td>
<td>Emphasizes access or availability of condoms</td>
</tr>
<tr>
<td>Condom proponents</td>
<td>Refers to personal implementation of condom use as a reduction/HIV-prevention strategy and includes also explicit statements of increased condom use</td>
</tr>
<tr>
<td>Condoms continued</td>
<td>Statements that condom use is continued or that precautions were taken prior to trial enrollment and that this behavior is continued</td>
</tr>
<tr>
<td>Condoms intent</td>
<td>Individual expresses intent to begin using condoms</td>
</tr>
<tr>
<td>Condoms not used</td>
<td>Condoms not used or used very infrequently</td>
</tr>
<tr>
<td>Disclose behavior</td>
<td>Participant hesitates or refrains from telling counselors/staff about risky behaviors</td>
</tr>
<tr>
<td>Drug sex connection</td>
<td>Descriptions of how sexual behavior, especially risky behavior, is influenced by, or related to, drug and alcohol use</td>
</tr>
<tr>
<td>Feel good</td>
<td>Statements that behavioral and knowledge changes have resulted in the individual’s feeling good, emotionally or physically</td>
</tr>
<tr>
<td>Gender preference</td>
<td>Expression of preference for a particular gender of counselor or potential problems associated with a specific gender of counselor</td>
</tr>
<tr>
<td>Get HIV tested</td>
<td>Perceptions or attitudes regarding availability and accessibility of HIV tests, including the process of becoming HIV-testing consumers</td>
</tr>
</tbody>
</table>

(continued)
we were not interested in determining the weight of each code for each respondent (i.e., frequency of application). Instead, we wanted to focus on examining the presence of codes across interviews. This approach obviates potential bias that would occur, for example, if a code were repeatedly applied at a high rate to only one respondent’s text.

Using AnSWR’s Boolean search functions, we selected all coding associated with our structural code of interest, RRC Experience. AnSWR does not include a cluster analysis function, so we decided to use ANTHROPAC. The binary matrix was saved as an ASCII-delimited file with a .DAT extension, subsequently opened in Windows Notepad and edited into a DL format suitable for ANTHROPAC.1 Once we had imported the data into ANTHROPAC, we converted the respondent-by-code (35 × 46) binary matrix into a one-mode, 46 × 46, code-by-code, symmetric, binary matrix and calculated similarities using cross-product matching (i.e., sum total of co-occurrences). We deleted the structural code from the code-by-code matrix (because the context had already been established and applied to all thirty-five respondents’ text) and ran a complete linkage hierarchical cluster analysis.

Hierarchical cluster analysis is an agglomerative methodology that identifies clusters of observations in a data set (Aldenderfer and Blashfield 1984). The three most common algorithms for clustering are single linkage, average linkage, and complete linkage. We used the complete linkage method, in which the farthest pair of observations is used to compute distances, because it usually performs well when the objects tend to form naturally distinct groupings. In addition, complete linkage cluster analysis requires no a priori information about the samples, so the tree reflects the largest differences and similarities among the samples.

Cluster analysis produces a relatively easy-to-read output based on the relationship between codes as they are applied to the raw data and on the frequency with which they co-occur. Such a method is not only data driven but also allows a researcher to see patterns in large data sets with a large number of codes. Moreover, a refined context can be created by selecting only the codes that are applied in concert with a particular structural code. Systematically retrieving and organizing data in this way give additional meaning to the analysis and provide analytical logic.

Figure 1 depicts the cluster tree diagram of men’s risk reduction counseling experiences within the context of HIV vaccine trial participation. The figure illustrates broad clusters that break down into several smaller clusters. In our 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. A cluster groups together particular semantic elements (i.e., codes). 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. As such, for our analysis, the levels of the clusters at the far left of the figure represent the highest number of distinct respondents associated with a given theme or cluster.

The cluster tree diagram provides a starting point from which to develop a more complete interpretation of the data. The most salient codes appearing in the highest level cluster, and consistent with the AnSWR-generated salience report, are No Hang Ups and Comfort Good. We applied the first of these codes to statements that represented a respondent’s ease or comfort in disclosing to a clinical nurse or counselor intimate details about his sexual risk-taking behavior. We applied the code Comfort Good when respondents directly or indirectly expressed that they had felt comfortable during the risk-reduction counseling. The association between these two concepts confirms what we would expect to be a logical and intuitive relationship. What is highly useful, however, is the guidance that this cluster provides for interpreting information in a larger framework.

As shown in Figure 1, the coded data break up into three clusters. Two of these clusters further split into subclusters. To determine the cut-off level for defining a cluster, we compared the cluster tree diagram to the salience

---

**Table 1 (continued)**

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safer external</td>
<td>Statements that address external factors attributed to an increase or maintenance of safe sex behavior practices</td>
</tr>
<tr>
<td>Safer sex</td>
<td>General references to safer sexual practices that do not specifically refer to partner numbers or condom use</td>
</tr>
<tr>
<td>Services referrals</td>
<td>Reference is made to services or referrals that are provided by trial staff, including counselors</td>
</tr>
<tr>
<td>Significant others</td>
<td>Statement that behavioral changes were attributable to significant others, such as children, grandchildren, parents, partners, and so forth</td>
</tr>
<tr>
<td>Single partner</td>
<td>Reference made to having a single sexual partner</td>
</tr>
<tr>
<td>Staff improvements</td>
<td>Suggestions for improving the staffing/counseling aspects, such as more consistency in counselor assignment, staff commitment, or counselor attributes</td>
</tr>
<tr>
<td>Study related</td>
<td>Emphasis is on influence or impact of study on participant’s behavior</td>
</tr>
</tbody>
</table>

NOTE: RRC = risk-reduction counseling.
report. At level 4, we found that more than 20% of respondents provided information associated with a code contained within a cluster. The tree diagram displays notable breaks in clusters for codes with salience with ≤ 20%, hence supporting the benchmark we chose.

Anchoring our discussion in high points of the tree diagram, we can move across high-level clusters and fill in the details along the way: We add relevant quotes from text, refer to the literature, and confirm our interpretation with saliency and frequency tables. The resultant story, grounded in the data and given a relatively objective structure, is summarized in the following paragraphs. For pedagogical purposes, we distinguish clusters from subclusters (themes, subthemes), examine the co-occurrence between codes within these clusters to identify core elements (codes x and y), and then review how information associated with these codes is specifically articulated (text). By doing so, we demonstrate how to create an interwoven narrative that is structured by the configuration of clusters.

Cluster 1: Rapport Theme

Of the men interviewed, most reported having no problem with discussing their sexual behavior and were comfortable with the risk-reduction counseling (subcluster A: No Hang Ups and Comfort Good). Many indicated that they were accustomed to discussing their sexual behavior with others and/or being part of a research project (text). Respondents had difficulty distinguishing between the risk-assessment data-collection process and the HIV risk-reduction counseling. Risk-assessment questions were frequently connected to an increased self-awareness of risk behavior (subcluster B: Assessment of Risks and Aware Know). Explicit vocalization of risk behavior was reported to be an “eye-opening” experience (text). Overall, a positive trial experience related to having a good counselor (subcluster C: Positive Experience and Good Counselor), someone with whom the participant felt a certain degree of rapport (text). Two commonly expressed characteristics of a good counselor were sincerity and a nonjudgmental attitude (text).

Cluster 2: Balancing Risks Theme

Participants explained that they had established acceptable boundaries for sexual behavior, but disclosing risky or unconventional sexual practices, such as not using condoms, was uncomfortable (subcluster A: Acceptable Risks, Comfort Low, and Condoms Not Used). This discomfort was often manifested in the form of guilt or embarrassment (text). Participants indicated that an enjoyable and healthy sex life was important (text) but that it had to be balanced with safe sex behaviors, such as the use of condoms in situa-
Cluster 3: Risk Homeostasis Theme

Respondents reported that patterns in their sexual behavior had been well established prior to volunteering for the trial and that no or little change in their sexual practices had occurred while participating in the trial (subcluster A: Habits Fixed and No Behavior Change). Having received repeated HIV education over the course of the epidemic (text), many perceived that the counseling information contained "the same old messages" (subcluster B: Old Messages). Their relatively fixed repertoires and high level of HIV knowledge were not immutable, however (text), but any changes would be dependent on new developments in HIV research (text).

CONCLUSION

Cluster analysis and our in-depth familiarity with the raw data allowed us to construct a structured narrative that is grounded in data. Of course, in an actual analysis, we would provide more details in our interpretation and flesh out examples as appropriate. We would also support our narrative with verbatim quotations and numeric output such as frequencies and salience, referring to relevant publications in the process. Our brief narrative above is intended to give the reader an idea of how cluster analysis can help frame thematic analysis and interpretation. High-level clusters can serve as conceptual starting points that create a framework for discussion, often one of the most difficult parts of the interpretive process. This framework can be filled with details derived from the text.

This type of analysis can, of course, be taken a step further. For example, matrices generated for different structural codes could be compared to determine if a single (i.e., identical set of codes are contained in each domain of inquiry) or multiple (i.e., unique codes for each domain of inquiry with potential overlap between domains) analysis was necessary. One could also produce an overall cluster analysis, not defined by any structural code, to create relatively context-free clusters. Ultimately, one's approach will depend on the type of data, the granularity of coding, and the research objectives.

Cluster analysis can be both a useful and a powerful tool, but its use must be tempered with common sense and in-depth knowledge of the raw data. Interpretation is only as good as the procedures that precede it (i.e., code development and application), and applying structure to an unfamiliar data set will have little meaning for even the most seasoned researcher. Used properly and solidly grounded in the data, cluster analysis can help provide meaningful structure to QDA and thematic interpretation.

NOTE

1. There are a number of qualitative data analysis software programs, such as NUD*IST, TextSmart, WordStat, and Hamlet, that do include this function. Data generated in qualitative data analysis programs can also be imported into other packages such as UCINET, SPSS, SAS, and CATFAC.

REFERENCES

MacQueen, K. M., E. McLellan, D. S. Metzger, S. Kegelies, R. Strauss, R. Scotti, L. Blanchard,
Miles, M. B., and M. A. Huberman. 1994. Qualitative data analysis: An expanded sourcebook.
Osgood, C. 1959. The representational model and relevant research methods. In Trends in content
Ryan, G. 1999. Measuring typicity of text: Using multiple coders for more than just reliability
for Word-based records. Atlanta, GA: Centers for Disease Control and Prevention.

GREG GUEST, Ph.D., is a contracted health care analyst at the Centers for Disease Control
and Prevention in Atlanta, Georgia. He is an applied anthropologist whose research interests include HIV risk behavior, political ecology of health, and integrating qualitative and quantitative methods. His methodological publications include "Using Gutman Sealing to Rank Wealth: Integrating Qualitative and Quantitative Data" (2000, Field Methods) and "An Ethnographic Approach to Design" (with J. Blomberg and M. Burrell, in Handbook for Human-Computer Interaction, 2002, Lawrence Erlbaum).

ELEANOR McLELLAN, M.A. is a health scientist with the Centers for Disease Control
and Prevention in Atlanta, Georgia. Her research interests and experiences include applied anthropology, particularly in HIV prevention and clinical trials, community development, gender issues, ethnic relations, and research methodology. Her work has involved the development of AnSWR: Analysis Software for Word-Based Records. Her publications include "Beyond the Qualitative Interview: Data Preparation and Transcription" (with K. M. MacQueen and J. L. Nesvig, in press, Field Methods), "Codebook Development for Team-Based Qualitative Analysis" (with K. M. MacQueen, K. Kay,