Intercoder Agreement in Analysis of Responses to Open-Ended Interview Questions: Examples from Tuberculosis Research

James W. Carey
Centers for Disease Control and Prevention (CDC), Atlanta
Mark Morgan
1310 N. Lake Drive, Canton, GA 30115
Margaret J. Oxtoby
New York State Department of Health

Cultural Anthropology Methods 8(3):1-5

Introduction

Social scientists commonly conduct surveys to learn about beliefs, attitudes, reported behaviors, or experiences prevalent in a population. These interviews often involve the use of structured questions with closed-ended responses (e.g., true/false, agree/disagree, or multiple choice). Open-ended questions can also yield useful information, especially when researchers need to explore complex issues that do not have a finite or predetermined set of responses.

A number of researchers have tried to develop systematic methods for analyzing written narratives generated through open-ended surveys or other qualitative data collection techniques (e.g., Bernard 1994; Carey 1994a, 1994b; Gorden 1992; Miles and Huberman 1994; Patton 1990). Skeptics often criticize qualitative data analysis procedures as being overly subjective. A researcher's failure to provide a replicable analysis may jeopardize the scientific credibility of the results. To overcome this difficulty, investigators should take explicit steps to ensure strong intercoder agreement in qualitative data analysis.

Recently, we completed a study of beliefs and treatment concerns about tuberculosis (TB) among newly arrived Vietnamese refugees in two counties in New York State. Data were collected by two trained Vietnamese interviewers who spoke English as a second language, and they asked each adult respondent (age 18 ≥ years, N=51) 32 open-ended questions (Carey, Oxtoby, and Carloni 1994). Question topics included knowledge and beliefs about TB symptoms and causes, as well as beliefs about susceptibility to the disease, prognosis for those who contract the disease, skin testing procedures, and prevention and treatment methods. Respondents were also asked how they thought TB affected one's ability to work and what they thought the reactions (e.g., fear, rejection) might be from family, friends, and community members.

In addition, interviewers collected data about the respondents' demographic, socioeconomic, and educational backgrounds. The purpose of the research was to identify health education needs, since misconceptions and concerns about TB can interfere with successful completion of treatment. The study was developed as one step in improving the quality and effectiveness of TB health care services among Vietnamese refugees in New York State and elsewhere in the United States.

This paper describes the methods we used to strengthen and measure the replicability of the analysis of the open-ended interview responses. Other investigators may find these techniques applicable in their work. The findings and recommendations from the study are presented elsewhere (Carey et al. 1995).

Data File Organization

Interviews were conducted in Vietnamese. The interviewers recorded notes and then used the notes to write expanded English summaries of each respondent's answers to the questions. These summaries were put into 32 separate ASCII files, one question per file, using a word processor. Each file thus contained 51 passages representing all the respondents' answers to a single question, and these were marked with identification numbers and delimited with brackets.

Individual answers varied in length, complexity, content, and the extent to which they agreed with conventional medical knowledge about TB. For example, the third question asked during the interviews was: "What do you think causes tuberculosis?" The biomedical answer is that tuberculosis is caused by *Mycobacterium tuberculosis*, a bacterial pathogen first discovered by Robert Koch in the late 1800s. The following quotes from the cause question file show responses from the first two respondents. Text passages for the other 49 individuals followed the first two, and similarly structured files contained the responses for the other 31 questions:

> [.id pat01 TB sometimes caused by working too much, and too hard (overworking). For example, some Vietnamese that work beyond their strength like farmers, factory workers, are working all day (some work 12, 14 hours per day, 7 days per week) and when come home and very little and also no nutritional food. When you overwork and don't have enough calories in your body, it is very easy to get all kinds of germs to enter your body because your body's immune system is not strong enough. Also, they may acquire the TB germ from someone that has TB. Some cause by smoking too much, or drinking. The amount of nicotine and alcohol that you take into your body will be very harmful for lungs (if you have been smoking and drink a lot).]
>
> [.id pat02 I was told that everyone does have the Koch virus in his/her body, and if one is overworked without proper nutrition can get tuberculosis.]

The data were organized in this way to facilitate coding and analysis using Tally (Bowyer 1991). See Trotter (1993) for a review of Tally. Tally is one of many software options available for analysis of qualitative data. Even though Tally and most other currently available packages do not help researchers compute agreement indices, we chose it because it was easy to use, met most of our other analysis needs, and the coding procedures are compatible with the more advanced AnSWR software package that the Centers for Disease Control and Prevention (CDC) is currently developing (Milstein and MacQueen 1994). The intercoder agreement methods described in this paper could be modified for use with other software, and AnSWR will include a variety of methods for assessing intercoder agreement.

Coding Procedures

The coding of qualitative data entails assigning unique labels to text passages that contain references to specific categories of information (Bernard 1994; Gorden 1992; Miles and Huberman 1994). For our study, we needed to develop and assign a list of codes that corresponded to each separate TB-related belief held by the respondents. Prior to the research, we had little idea what Vietnamese refugees might know or believe about tuberculosis. If we had known the range of Vietnamese TB beliefs ahead of time, it would have been possible to design an interview instrument consisting of structured response questions in place of the open-ended items.

Creation of the code list was an inductive task, based on what respondents said. We began by reading responses from the interviews and then compiled a code book containing a list of mnemonic codes along with their definitions. For example, the letter sequence "CAUSESMK" referred to the belief that TB could be caused by smoking. As each new idea or belief was encountered, it was added to the code book.

Gorden (1992:181) has stated that a useful set of codes should be all-inclusive and mutually exclusive. Our final code book contained 171 unique codes and their definitions, each corresponding to various TB beliefs stated at least once by one or more of the 51 respondents. Table 1 shows examples from our code book.

Tally, and other comparable software, helps the researcher attach codes to segments of text such as a word, phrase, sentence, or paragraph (Miles and Huberman 1994; Milstein and MacQueen 1994; Weitzman and Miles 1995). In our study, a segment was equivalent to each of a respondent's responses.[1] This yielded 1632 text segments, corresponding to the 32 question-specific response summaries from the 51 individuals.

Since many responses contained multiple beliefs, the number of codes assigned to each passage varied. A specific code could be used only once per response. Referring once more to our earlier example describing the first respondent's beliefs about TB causation, we assigned seven different codes. These included overwork in hard manual labor (CAUSEWRK), poor nutrition (CAUSENUT), infection by microorganisms (CAUSEGRM), weakened immune systems (CAUSEIMM), exposure to someone with contagious or infectious TB disease (CAUSECON), smoking (CAUSESMK), and alcohol consumption (CAUSEDRK).

The passage for the second respondent received three codes: CAUSEWRK, CAUSENUT, and a new code, CAUSEKOC, which was used to label the statement concerning the Koch virus.

Intercoder Agreement Methods

Once a working draft of the code book was developed, our next step was to ensure that different coders could independently replicate each other's work using the same instructions. To pretest the code book and estimate final intercoder agreement, we selected 320 passages of text comprising the responses given by the first 10 individuals in the sample for each of the 32 questions. This represented nearly 20% of the 1632 response summaries.[1]

Miles and Huberman (1994:64) and Gorden (1992:185) emphasize the importance of pretesting and revising code books, because initial coding instructions often yield poor agreement. In our study, two coders independently coded the 320 text segments two times. The purpose of the first coding comparison was to pretest and remedy problems with the code book. Once these problems were identified and fixed, the same two coders used the amended code book to recode the same 320 passages of text. This let us estimate the final, or achieved, degree of intercoder agreement.

We used two methods to quantify the degree of agreement during the code book pretest and final intercoder agreement assessment phases. In the first technique, we compared the sets of codes that each coder assigned to each of the 320 text passages. A response was considered to be coded the same only if both coders used the identical set of codes. For example, if one coder assigned the CAUSESMK, CAUSENUT, and CAUSEDRK codes to a response, the other coder had to assign the same three codes in order for there to be agreement. Presence of one or more disagreements, such as not assigning one of these codes or assigning a fourth code, was counted as a coding discrepancy.

Using this method, comparison of the code book pretest results showed that only 144 (45.0%) out of the 320 responses were coded the same by both coders. This poor level of replicability indicated a need to further refine the code book.

By discussing the reasons for their disagreements, the two coders were able to identify and correct problems with the code book. The reasons for the discrepancies included problems such as redundant codes for the same belief, vague code definitions, lack of mutual exclusivity between codes, or lack of shared understanding in the procedures for using specific codes.

After resolving the unclear parts of the code book, the two coders recoded the same 320 responses a final time using the revised code book. The final level of agreement between the two coders showed substantial improvement. Using the amended code book, the coders were in complete agreement for 282 (88.1%) of the 320 responses. There were only 38 passages (11.9%) with one or more discrepancies. In sum, code book clarifications from the pretest made it

possible to nearly double the final level of agreement.

The second intercoder agreement method involved examination of agreement in how each separate code was used by the two coders. To do this, we constructed a 2-by-2 contingency table showing the presence or absence of the belief as judged by both coders across the 320 text passages. There were two agreement, or concordant cells: the upper left cell indicated the number of times that both coders assigned the code, and the lower right cell contained the number of times both coders did not assign the code to a response.

The two disagreement, or discordant cells showed the number of times a code was assigned by one coder but not the other. A separate table was computed for 152 of the 171 codes listed in the final version of the code book. The other 19 codes pertained to beliefs not expressed by the first ten respondents (and that therefore did not apply to the 320 text passages used for estimating agreement), or for which reliability statistics could not be computed due to zeros in the 2-by-2 table margins. The process was repeated twice, once during the code book pretest phase and again as a means to estimate final intercoder agreement.

A deceptively simple measure of agreement for how a code is assigned might be the proportion of times that the two raters agreed (i.e., the sum of the two agreement cells, divided by 320). However, simple proportions do not account for the possibility that coders might agree due to chance (Gorden 1992). Therefore, the proportion of agreement is a biased overestimate of the true level of agreement. To correct for this, we used the kappa statistic.[2]

During the code book pretest, we judged that a kappa <0.90 indicated a problem with agreement in the way a code was being used. Examining the pretest kappa values helped focus remedial attention on those parts of the code book with the poorest agreement. Recalculation of all 152 kappas following amendments to the code book indicated that considerable improvements in agreement had been achieved.

Table 2 shows the number of codes and corresponding final kappa values.

Final kappa values indicated that the two coders used 126 (82.9%) of the 152 codes in exactly the same way (kappa = 1.0) across all 320 text passages. Kappa values of 0.90 or greater were achieved for 135 (88.8%) of the codes. Only 17 (11.2%) of the 152 codes had final kappa values $\leq$0.89. As a way to ensure consistency in subsequent analysis steps, the senior author resolved any remaining intercoder discrepancies in the 320 text passages, and assigned codes for the rest of the data (Carey et al. 1996).

Discussion

In our study of TB beliefs among recent Vietnamese refugees, we used an array of techniques to strengthen and measure agreement in the coding of responses to open-ended questions. We believe that other qualitative researchers should increase their efforts to ensure coding replicability. Critics of qualitative research are correct in pointing out that many studies do not adequately address this issue. Some qualitative researchers even claim that one should not attempt to ensure replicability in qualitative data analysis. However, subjective biases introduced into the coding and analysis of qualitative data have adverse consequences.

In basic research, coding biases can lead to inappropriate theoretical conclusions, and in applied research the problem can lead to ineffective or harmful recommendations and interventions. Although it is not an absolute guarantee against coding bias, failure to make any attempt to strengthen and measure agreement can undermine the credibility, validity, and utility of qualitative research findings

At least five aspects of our approach may be generalizable to other studies employing open-ended interview questions. First, careful research design is an important prerequisite to achieving reliable qualitative analyses. Development of a systematic data management and analysis plan prior to data collection can help one avoid many difficulties common in qualitative data analysis.

Second, as has been previously emphasized in the literature, development of a good set of qualitative data coding

instructions is an inductive process requiring multiple iterations. By pretesting the code book, we gained an opportunity to identify and rectify unreliable aspects of our coding procedures. Once code book problems were fixed, we attained a final, or achieved, level of agreement of over 88%.

Miles and Huberman (1994:64) suggest that final intercoder agreement in qualitative data analysis should approach or exceed 90%. Our experiences highlight the value and necessity of systematic testing and revision of qualitative data coding procedures. Before we conducted our pretest, we erroneously assumed that the code book was comprehensive and clear. If we had not made the effort to identify and correct problems, we could have unknowingly generated invalid findings and faulty conclusions.

Third, initial organization of the database greatly facilitated intercoder comparisons and subsequent analysis. By placing all of a question's answers into separate files, each file had a similar physical structure. This automatically subdivided the text into standardized segments for the two coders to use, and eliminated potential disagreements caused by differences in segmentation. Other text data sets may not have such natural or easily identifiable segments, which may require intercoder comparison of segmentation judgments.

Fourth, because of the volume of data and the time constraints, we did not attempt to ascertain agreement for the whole database. Instead, we chose the responses from approximately the first 20% of the respondents to act as a standard data subset for pretest and final agreement estimation.

Fifth, other researchers may find it useful to measure intercoder agreement using multiple methods. In our study, we first compared the sets of codes assigned by the two coders to the response segments. This method was relatively fast and easy. Another advantage of this approach is that the researchers are comparing overall interpretation of text segments rather than simply looking at agreement in using individual codes.

However, examination of code-specific agreement levels using the pretest kappa statistics helped us to pinpoint precisely which codes were giving us the greatest problems. Examination of code-specific agreement using the kappa statistic demonstrated that we had developed a strongly replicable set of coding instructions.

The disadvantage of this method is that computation of the kappa statistics is more time consuming. Researchers short on time may wish to use only the first method, since the final agreement results using the two methods are similar.

Acknowledgments

Many individuals contributed to this research and paper, among them are Nancy Binkin, Ed Carloni, Jeannette Castiglione, Ken Castro, George DiFerdinando, Larry Geiter, John Grabau, Michael Hennessy, Van Huynh, Marva Jeffery, Ann Lanner, Kate MacQueen, Bess Miller, Bobby Milstein, Lien Pham Nguyen, Lee Paul, Arthur Rubel, Richard Sessler, Phil Smith, Esther Sumartojo, and Neff Walker. Special thanks go to the Vietnamese people who participated in the study. Institutional assistance was provided by the New York State Health Department's Bureau of Tuberculosis Control and Refugee Health Programs, the Broome County Department of Health and Refugee Assistance Program, the Onieda County Department of Health and the Mohawk Valley Resource Center for Refugees, and the Council of State and Territorial Epidemiologists. Funding and additional institutional support was provided by the Division of Tuberculosis Elimination at the Centers for Disease Control and Prevention.

Notes

1. Text passages from the first 10 respondents were selected because their data became available early in the project's data collection phase. This allowed us to begin prompt development and testing of the code book. Although it would have delayed onset and completion of the analysis, an alternative procedure would have been to select a random subset text passages from all 51 respondents after all the interviews were completed. A mean of 2.4 codes were used per response segment, with a range of 1 to 13 codes per response.

2. Kappa is a measure of the amount of agreement between two coders after statistically adjusting for agreement due to chance. Total agreement between two coders yields a kappa=1.00. Any disagreement produces a value <1.00, with lower values indicating larger discrepancies. Kappa takes negative values when there is less agreement than expected by chance alone (Fleiss 1981).

References

Bernard, H. R. 1994. *Research Methods in Anthropology: Qualitative and Quantitative Approaches.* 2d edition. Thousand Oaks, CA: Sage Publications.

Bowyer, J. W. 1991. *Tally: A Text Analysis Tool for the Liberal Arts.* Dubuque, IA: William C. Brown Publishers.

Carey, J. W. 1994a. Methods for Analyzing Responses to Open-Ended Survey Questions. *TB Notes* Summer:13–14. Atlanta: Centers for Disease Control and Prevention, Division of Tuberculosis Elimination.

Carey, J. W. 1994b. Improving International HIV Program Planning: Systematic Interview Methods in the Context of Programmatic Needs Assessment. Masters thesis in Public Health, Emory University, Atlanta.

Carey, J. W., M. Oxtoby, and E. Carloni 1994. Prevention of Tuberculosis among Vietnamese in New York State: Interviewer Skills Training Workshop. Unpublished instructional manual, pp. 1–51.

Carey, J. W., M. J. Oxtoby, L. P. Nguyen, V. Huynh, M. Morgan, and M. Jeffery 1995. Tuberculosis Beliefs among Recent Vietnamese Refugees in New York State: Implications for Improving Patient Education and Therapeutic Adherence. Manuscript under revision for submission to *Public Health Reports.*

Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions.* 2d edition. New York: John Wiley.

Gorden, R. 1992. *Basic Interviewing Skills.* Itasca, IL: F. E. Peacock.

Miles, M. B. and A. M. Huberman 1994. *Qualitative Data Analysis.* 2d edition. Thousand Oaks, CA: Sage Publications.

Milstein, B. and K. MacQueen 1994. AnSWR on the horizon. *Practicing Anthropology 16*(3):34–35.

Patton, M. Q. 1990. *Qualitative Evaluation and Research Methods.* Newbury Park, CA: Sage Publications.

Trotter, R.T. 1993. Review of TALLY 3.0. *CAM* 5(2):10-12.

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