Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research

Article *in* Field Methods · August 2004

DOI: 10.1177/1525822X04266540

Daniel Hruschka
Arizona State University
134 PUBLICATIONS **2,695** CITATIONS
SEE PROFILE

Richard Jenkins
National Institute on Drug Abuse
66 PUBLICATIONS **2,548** CITATIONS
SEE PROFILE

James W Carey
Centers for Disease Control and Prevention
48 PUBLICATIONS **2,226** CITATIONS
SEE PROFILE

Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research

DANIEL J. HRUSCHKA
Centers for Disease Control and Prevention, Atlanta, Georgia
Emory University
DEBORAH SCHWARTZ
DAPHNE COBB ST. JOHN
ERIN PICONE-DECARO
RICHARD A. JENKINS
JAMES W. CAREY
Centers for Disease Control and Prevention, Atlanta, Georgia

Analysis of text from open-ended interviews has become an important research tool in numerous fields, including business, education, and health research. Coding is an essential part of such analysis, but questions of quality control in the coding process have generally received little attention. This article examines the text coding process applied to three HIV-related studies conducted with the Centers for Disease Control and Prevention considering populations in the United States and Zimbabwe. Based on experience coding data from these studies, we conclude that (1) a team of coders will initially produce very different codings, but (2) it is possible, through a process of codebook revision and recoding, to establish strong levels of intercoder reliability (e.g., most codes with kappa ≥ 0.8). Furthermore, steps can be taken to improve initially poor intercoder reliability and to reduce the number of iterations required to generate stronger intercoder reliability.

Keywords: *intercoder agreement; interrater agreement; open-ended; qualitative data; reliability*

In the past decade, qualitative research methods have attracted a great deal of attention in business (Sykes 1991), consumer research (Kolbe and Burnett 1991), public health (Mantell, DiVittis, and Auerbach 1997), nursing (Field and Morse 1985; Appleton 1995), health care research (Fitzpatrick and

We gratefully acknowledge the insightful comments provided by Ron Stall, Eleanor McLellan, John Williamson, the editor of Field Methods*, and several anonymous reviewers on earlier drafts of this paper.*

Field Methods, Vol. 16, No. 3, August 2004 307–331
DOI: 10.1177/1525822X04266540

Boulton 1996), social work (Drisko 1997), and health fields in general (Mays and Pope 1995). Text from transcripts and interviews constitutes the bulk of data used in this research, but a range of object-oriented forms (visual images, videos, and audio segments) have also been considered. Although researchers have proposed general guidelines for analysis of the large amounts of data gathered in such forms (Field and Morse 1985; Weber 1990; Gorden 1992; Denzin and Lincoln 1994; Miles and Huberman 1994; Carey, Morgan, and Oxtoby 1996; MacQueen, McLelland, and Milstein 1998; Ryan and Bernard 2003), most treatments only briefly address specific questions of methodological importance, such as intercoder reliability (Gorden 1992:173–90; Miles and Huberman 1994:50–67; Carey, Morgan, and Oxtoby 1996; MacQueen, McLelland, and Milstein 1998).

Most approaches to qualitative data analysis involve the identification and coding of themes that appear in text passages (or other media segments). Coding entails (1) compiling a list of defined codes (the codebook) corresponding to themes observed in a text and (2) judging for each predetermined segment of text whether a specific code is present. Although this procedure is a standard in qualitative data analysis, assessing the degree to which coders can agree on codes (intercoder reliability) is a contested part of this process (Armstrong et al. 1997; Mays and Pope 2000). Some researchers argue that qualitative inquiry is a distinct paradigm and should not be judged by criteria, such as reliability, that are derived from “positivist” or “quantitative” traditions (Guba and Lincoln 1994; Madill, Jordan, and Shirley 2000). Others have expressed skepticism about the subjective nature of qualitative analysis and further question whether it is possible to generate reliable codings (Weinberger et al. 1998; for review, see Mays and Pope 1995). The common assumption that generating reliable codings of text is impossible, or at best of minor importance, manifests itself in the haphazard and unclear reporting of intercoder reliability in many qualitative research studies (for reviews, see Kolbe and Burnett 1991; Lombard, Snyder-Duch, and Campanella 2002). In contrast to these views, a third position holds that intercoder reliability is a useful concept in settings characterized by applied, multidisciplinary, or team-based work (Krippendorff 1980; Weber 1990; General Accounting Office [GAO] 1991; Gorden 1992; Miles and Huberman 1994; Carey, Morgan, and Oxtoby 1996; Armstrong et al. 1997; Boyatzis 1998; MacQueen, McLelland, and Milstein 1998). This view is informed partly by research in cognitive science and decision making that has shown that there are limits to the human ability to process the kind of complex information often amassed by qualitative research (Simon 1981; Klahr and Kotovsky 1989). When making judgments based on complex data, for example, people often use intuitive heuristics that may introduce bias or random error (Kahneman, Slovic, and

Tversky 1982). Specifically, the high degree of inference required to categorize types of open-ended responses can lead to initially low agreement between coders (Hagelin 1999). Establishing intercoder reliability is an attempt to reduce the error and bias generated when individuals (perhaps unconsciously) take shortcuts when processing the voluminous amount of text-based data generated by qualitative inquiry.

The applied and multidisciplinary requirements of qualitative research at the Centers for Disease Control and Prevention (CDC) have made intercoder reliability an important criterion for assessing the quality of findings. CDC-supported research is only as valuable as its applicability to real world problems, and policy decisions based on unreliable findings risk wasting resources and endangering public health (Carey, Morgan, and Oxtoby 1996). Furthermore, findings in public health often are presented to multidisciplinary audiences, and to communicate credibly to persons with diverse theoretical and methodological backgrounds, it is necessary to clearly describe how conclusions have been derived from the data. The logic of reliability, in general, and intercoder reliability, in particular, is recognized across a variety of disciplines as a measure of the quality of one stage in the research process.

In this article, we describe one process developed at the CDC for reliably coding texts. We describe this coding process applied to three HIV-related qualitative studies (the Los Angeles Bathhouse Study; the Acceptability of Barrier Methods to Prevent HIV/AIDS in Zimbabwe, Africa, study; and the Hemophilia Study). Using studies that vary in design, purpose, and type of qualitative data, we show that (1) coders initially generate very different codings of text, (2) intercoder agreement improves substantially following a systematic process to revise and test the codebook, and (3) steps can be taken to improve initial intercoder agreement and to reduce the number of coding rounds needed to reach acceptable levels of intercoder agreement. These procedures were developed for HIV-prevention research, but they should be applicable to a broader range of subjects.

WHY INTERCODER RELIABILITY?

In psychometric literature, many questions about reliability are concerned with asking, "if people were tested twice, would the two score reports agree?" (Cronbach 1990:191). More general, classical reliability theory is concerned with assessing to what degree a measuring device introduces random error into the measurements of a unit of observation (Nunnally and Bernstein 1994). For example, fifty people guessing at the weight of an individual will generate a wider range of responses than fifty people painstakingly weighing the person by putting him or her on a scale. Because a person's weight is unlikely to change between weighings and guesses, the added variation in estimates introduced by guessing versus weighing would be considered added random measurement error. Because weighing introduces less random measurement error than guessing, we would say that it is a more reliable method of estimation. Of particular note here is that we can estimate the added random error without even knowing the true weight of the individual. We only need to look at the variation in a set of measurements on the same object. Although high levels of reliability in a measure are not sufficient for accurate measurement (a scale may consistently underweigh an individual), low reliability does limit the accuracy of measurements. For this reason, ensuring reliability is a necessary (but not sufficient) step for drawing accurate conclusions.

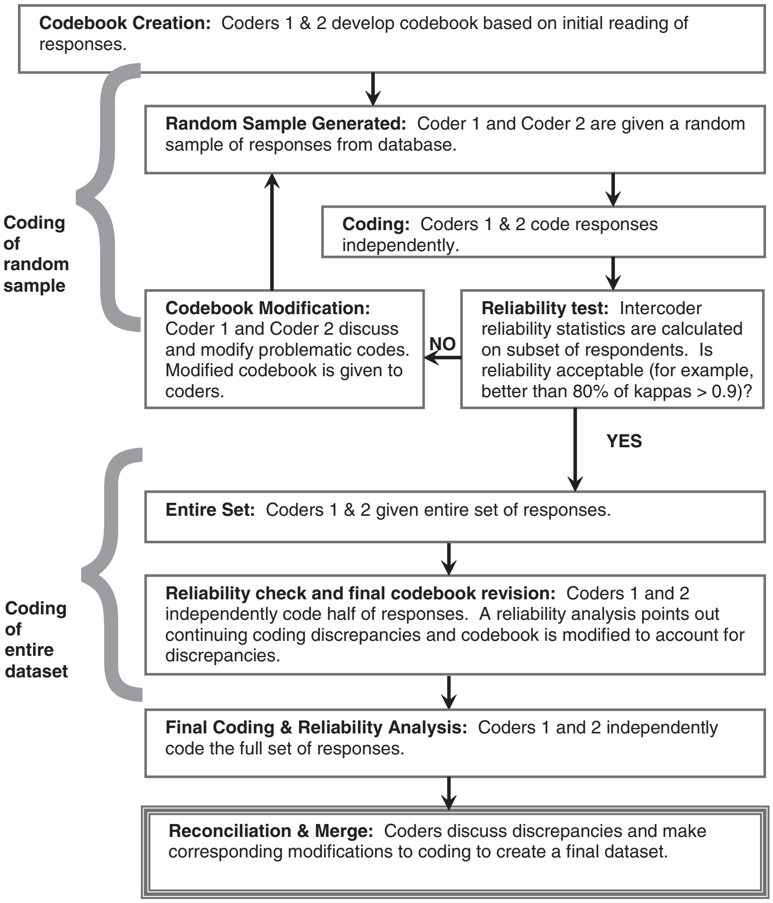
In assessing the reliability of text coding, a text segment is the unit of observation and each coding of that segment is a measurement. Once it has been written and divided into codable segments (words, sentences, paragraphs, responses), the text derived from an interview does not change. However, different coders may vary in their interpretation of the text's content. A systematic coding process, consistently used by each coder, should be more reliable compared with a process where each coder uses his or her own idiosyncratic methods (Miles and Huberman 1994; MacQueen, McLelland, and Milstein 1998; Boyatzis 1998). Intercoder reliability assesses the degree to which codings of text by multiple coders are similar. With intercoder reliability, the more coders (using the same codebook) agree on the coding of a text, the more we can consider the codebook a reliable instrument (i.e., one that facilitates intercoder reliability) for measuring the thematic content of a specific body of texts.

The Intercoder Reliability Process

To achieve acceptable levels of reliability, the process of coding text entails several steps: segmentation of text, codebook creation, coding, assessment of reliability, codebook modification, and final coding—with coding, assessment of reliability, and codebook modification perhaps conducted several times in iteration (see Figure 1).

Segmentation of text. Codes are applied to meaningful units of text usually referred to as segments. It is therefore essential to segment the text before coding begins. The segments may represent individual words, sentences,

FIGURE 1
Process for Generating Intercoder Reliability


Codebook Creation: Coders 1 & 2 develop codebook based on initial reading of responses.
Random Sample Generated: Coder 1 and Coder 2 are given a random sample of responses from database.
Coding: Coders 1 & 2 code responses independently.
Coding of random sample
Codebook Modification: Coder 1 and Coder 2 discuss and modify problematic codes. Modified codebook is given to coders.
NO
Reliability test: Intercoder reliability statistics are calculated on subset of respondents. Is reliability acceptable (for example, better than 80% of kappas > 0.9)?
YES
Entire Set: Coders 1 & 2 given entire set of responses.
Coding of entire dataset
Reliability check and final codebook revision: Coders 1 and 2 independently code half of responses. A reliability analysis points out continuing coding discrepancies and codebook is modified to account for discrepancies.
Final Coding & Reliability Analysis: Coders 1 and 2 independently code the full set of responses.
Reconciliation & Merge: Coders discuss discrepancies and make corresponding modifications to coding to create a final dataset.

paragraphs, responses to individual questions, or entire interviews. The question of how to divide texts into codeable units has no simple solution (Krippendorff 1995). With the three studies presented here, however, responses to individual questions were brief and were counted as single units (generally one line to one page per question).

Codebook creation. To generate an initial draft codebook, a portion of the data (e.g., the set of responses to a specific question) is distributed to a team of coders (often two, but preferably more). Team members independently examine the responses and propose a set of themes. The team meets to compare proposed themes and to agree on an initial master list of codes that operationalize these themes, paying close attention to (1) how relevant the codes are to current study goals and (2) whether the code actually emerges in the text. For each code, the team derives a set of rules by which coders decide whether a specific unit of text is or is not an instance of that code. Specifically, MacQueen, McLelland, and Milstein (1998) and Boyatzis (1998:31) have discussed schemes for efficiently defining a code that provides inclusion and exclusion criteria to clarify what segments of text do and do not constitute an instance of that code. Although it is possible to have codes with multiple values (e.g., high, medium, low), the simplest codes are dichotomous, indicating only if it is present or absent from a specific text segment.

Coding. After the initial draft codebook is developed, the team begins an iterative process of coding, reliability assessment, codebook modification, and recoding. Each iteration will be called a *coding round*. First, a "lead" coder assembles the draft codebook and distributes a subset of the raw uncoded data to the team of coders. Optimally, this subset should be randomly chosen from the respondents in a sample, but this may not always be possible because of resource or time constraints or a limited number of responses (Carey, Morgan, and Oxtoby 1996). For example, in a study with 300 respondents, it may be possible to randomly select a sample of 60 (20%) responses to capture variation, while in a study with 30 respondents, it may be necessary to consider all responses to capture appropriate variation. Once given the responses, each team member independently codes them according to instructions included in the draft codebook. The team meets again to discuss problems with applying codes, code definitions, and inclusion/exclusion criteria and to evaluate intercoder reliability.

Assessing intercoder reliability. A number of statistics can assess to what degree a set of texts were consistently coded by different coders (Krippendorff 1980; Carey, Morgan, and Oxtoby 1996). The commonly used "coefficient of agreement" (Neumark-Sztainer and Story 1997; Wang, Lin, and Ing-Tau Kuo 1997; see Kolbe and Burnett [1991] for review), which measures the proportion of decisions where coders agree, can dramatically overestimate the true degree of intercoder reliability by not taking chance agreement into account. Therefore, we relied on Cohen's kappa (Cohen

1960), which prevents the inflation of reliability scores by correcting for chance agreement, although other statistics also satisfy these criteria (Cohen 1960; Banerjee et al. 1999; Potter and Levine-Donnerstein 1999). The kappa measure can range from 1 to negative values no less than –1, with 1 signaling *perfect agreement* and 0 indicating *agreement no better than chance* (Liebetrau 1983). In practice, negative values are rare and indicate observed levels of disagreement greater than one would expect by chance. Achievement of perfect agreement is difficult and often impractical given finite resource and time constraints. Several different taxonomies have been offered for interpreting kappa values that offer different criteria, although the criteria for identifying "excellent" or "almost perfect" agreement tend to be similar. Landis and Koch (1977) proposed the following convention: 0.81–1.00 = *almost perfect*; 0.61–0.80 = *substantial*; 0.41–0.60 = *moderate*; 0.21–0.40 = *fair*; 0.00–0.20 = *slight*; and < 0.00 = *poor*. Adapting Landis and Koch's work, Cicchetti (1994) proposed the following: 0.75–1.00 = *excellent*; 0.60–0.74 = *good*; 0.40–0.59 = *fair*; and < 0.40 = *poor*. Fleiss (1981) proposed similar criteria. Cicchetti's criteria consider reliability in terms of clinical applications rather than research; hence, the upper levels are somewhat more stringent. Miles and Huberman (1994) do not specify a particular intercoder measure, but they do suggest that intercoder reliability should approach 0.90, although the size and range of the coding scheme may not permit this. In the studies presented below, we used fairly stringent cutoffs at kappa ≥ 0.80 or 0.90, roughly between Cicchetti's and Miles and Huberman's criteria.

Codebook modification. If intercoder reliability is judged to be insufficient, then the team discusses problems with the code definitions and proposes clarifications. If changes are made, the lead coder revises the codebook, distributes another subset of the raw data to the team, and the coding process is repeated until sufficient intercoder agreement is achieved (Miles and Huberman 1994; Mantell, DiVittis, and Auerbach 1997:171; MacQueen, McLelland, and Milstein 1998).

Coding of entire dataset. The number of iterations (or coding rounds) required to reach acceptable levels of intercoder reliability may vary, in part depending on the complexity of the responses, interview format, or the codebook (Willms et al. 1990; Carey, Morgan, Oxtoby 1996). When sufficient intercoder agreement is achieved, the entire set of responses for the complete sample is coded according to the final codebook revision (if smaller subsets were used for codebook generation). Systematic intercoder

TABLE 1
Overview of Key Study Variables

	Zimbabwe: Clinic in Harare	*Hemophilia:* *U.S.–Wide Telephone Survey*	*Los Angeles: Two Bathhouses*
# of interviews coded	295	70	24
# of coders	2	2	2
Form of questions	Short open-ended questions within survey	Short open-ended questions	Semistructured
How responses were recorded	Researcher notes	Researcher notes	Transcripts from audiotapes
# of questions coded per interview	4	27	30
Length of response segments per question	1–5 lines	1–5 lines	1 line to 1 page
# of codes in codebook	70–80	200–205	330–340
# of potential codes per response	14–30	200–205 (all global)	5–20
# of codes actually used per individual response	0–5[a]	0–11[a]	1–6
# of global codes	2	All codes were global	4
Coding rounds	3,4,4,9 depending on question	2	8

a. Passages could be coded with no codes if the response (1) did not fit any of the available codes, (2) was irrelevant to the question, or (3) was not given.

reliability checks may be made at intermediate stages of this final coding (e.g., after 50% completion) to ensure continuing intercoder reliability. Finally, when the entire dataset is coded, the final intercoder reliability for each code should be assessed.

MATERIALS AND METHODS

The Studies

The studies examined in this article were conducted by the CDC and collaborating institutions during the past decade. In all cases, qualitative data were collected from semistructured interviews, transcribed, and input into CDC EZ-Text for data management and analysis (Carey et al. 1998). Initial codes were derived by identifying themes in a set of randomly selected text passages and generating code definitions for these themes. Then, coding consisted of deciding for every text segment and for each code whether the theme indexed by the code was present or absent in the segment. Despite these basic similarities, the studies differ in the length and complexity of responses, the degree of interviewer probing, the content of questions, the number of study participants, and the length and complexity of the interview protocol (see Table 1).

The hemophilia study (Adult Hemophilia Behavioral Intervention Evaluation Project [HBIEP]). The hemophilia study was designed to evaluate interventions to avert HIV transmission between HIV seropositive men with hemophilia and their uninfected female sex partners residing in various locations throughout the United States (Parsons et al. 1998). Semistructured interviews were conducted by telephone with a subsample (subsample $n = 70$ couples, HIV seropositive men and HIV seronegative female partners) of the larger evaluation study sample. The study sought to generate hypotheses, uncover themes, and develop a broad perspective on possible determinants of behaviors related to risk reduction of HIV transmission. Transcribed interviewer notes from twenty-seven open-ended questions typically ranged from one to five lines of text. As one of the team's first attempts at establishing reliable text codings, the HBIEP Study analysis was in many ways a pilot effort.

The Los Angeles bathhouse study (LA bathhouse). This study provided the formative research development for counseling and testing services to be offered in Los Angeles bathhouses serving men who have sex with men (MSM; Mutchler et al. In press). Designed as a small exploratory study, its purpose was to generate recommendations for training counselors, marketing the testing program, and determining where counseling services would be provided in each bathhouse. This study involved face-to-face interviews with twenty-four MSM patrons of two bathhouses in Los Angeles. The semistructured interview included thirty open-ended questions. Particular responses to single questions typically ranged from two lines to one page of verbatim transcriptions from audiotaped interviews. The questions addressed topics such as bathhouse visiting patterns and common activities in the bathhouse.

The Acceptability of Barrier Methods to Prevent HIV/STDs in Zimbabwe, Africa (Zimbabwe study). This longitudinal intervention study was designed to introduce and to assess the acceptability of various barrier methods among heterosexual women, as well as to determine patterns of contraceptive use in two phases (O'Leary et al. 2003). Open-ended responses about condom negotiation and the acceptability of different contraceptive methods were collected as part of detailed interviews during study visits. Responses from the four open-ended questions considered in this article were translated into English from Shona, and the transcribed responses ranged from one to ten lines of text each.

Applying the Intercoder Reliability Process

In each study presented here, the general process described earlier (Figure 1) was followed, but differences in data type and population size between the studies resulted in different specifications of the process. For example, coders for the LA bathhouse study coded all the respondents per coding round, whereas coders in the Zimbabwe study coded only a subset of 20% of respondents (60 out of nearly 300) per coding round. In addition, lessons learned from previous studies were transferred to later ones. Whereas a large set of global codes that applied to all questions was employed in the hemophilia study, the LA bathhouse and Zimbabwe studies used small sets of codes that were specific to each question.

The criteria for judging acceptable levels of intercoder reliability also changed between studies. The hemophilia study used kappa greater than 0.8 as a cutoff for acceptable intercoder reliability. As one of the team's first coding efforts, the hemophilia study involved only two coding rounds with the assessment of intercoder reliability. The LA bathhouse study required that 80% of codes have a kappa score greater than 0.9, whereas the Zimbabwe study required that 90% of codes have a kappa score greater than 0.9.

One practice that continued throughout these studies was the use of dichotomous (rather than ordinal or multilevel categorical) codes. Therefore, for every text segment and for every code, a coder decided whether the code applied or did not apply to the text segment. Several codes could be applied to any specific text segment, indicating that coders noted more than one theme in the text segment. Another practice that remained consistent between studies was the use of Cohen's kappa to assess the intercoder agreement (Cohen 1960). Intercoder reliability reports, including kappa statistics, were generated by the qualitative data analysis software used in all three studies (CDC EZ-Text; see Carey et al. 1998). Finally, all coders were contract or civil service employees of the CDC with a minimum of a bachelor's degree, with some possessing a master's degree. Most had been trained in anthropology, although training in epidemiology and geography was also present. Supervision was provided by master's and doctoral level CDC staff with degrees in anthropology and psychology and experience in HIV prevention research.

Examining Factors in the Intercoder Reliability Process

Several factors may affect the time and effort required to implement the intercoder reliability process. First, a larger sample size or a greater interview length may increase the complexity of coding tasks and therefore reduce levels of intercoder reliability. Second, variation in the content, length of response, or number of codes per question may affect the speed at which a team achieves an acceptable level of intercoder reliability. Third, variation in the clarity of the codebook and individual code definitions also may influence the reliability process.

Two considerations made analysis of factors across these studies difficult. First, the codebooks were structured differently across studies, with two studies having unique codes for each question (LA bathhouse and Zimbabwe) and one study having global codes that could be assigned at any point in the interview (Hemophilia). For this reason, we could only examine the effect of number of codes per question for data from the LA bathhouse and Zimbabwe studies. Furthermore, results of the intercoder reliability process at the level of question or code were not available from the LA bathhouse study. For this reason, the Zimbabwe study was the only study where we could make interquestion comparisons.

When comparing measures of intercoder reliability for specific codes, we will often refer to the kappas associated with codes. Although the kappa is a measure that depends not only the code but also on the coders, we have retained this usage in an attempt to simplify the presentation of results.

TABLE 2
Percentage of Codes[a] with Kappa ≥ 0.9 by Study, Question, and Coding Round

	Full Interview		*Zimbabwe Questions*			
Round	*Hemophilia*[b]	*Los Angeles Bathhouse*	*Q6A*	*Q4*	*Q6C*	*Q7A*
1	33.9	39.0	38.5	50.0	61.1	46.2
2	64.6	44.1	63.6	50.0	62.5	68.4
3		52.1	75.8	92.9	86.7	88.9
4		56.9	71.9		100.0	94.4
5		69.5	80.6			
6		73.3	77.4			
7		78.3	74.2			
8		82.6	83.9			
9		90.3				
Final[c]		85.4	96.7	83.3	100.0	93.8

a. Proportion of codes that were used in particular coding round.
b. Proportion of codes > 0.8.
c. Round 2 was the final round for hemophilia.

RESULTS

Low Initial Intercoder Reliability

Regardless of study or question, the first round of qualitative coding generated low levels of intercoder reliability (see Table 2). In the Hemophilia study, 32.9% (80/243) of codes had kappa ≥ 0.8 with an intercoder assessment in the first round. In the LA bathhouse study, only 39.0% of codes had kappa ≥ 0.9 in the first iteration. In the Zimbabwe study, the responses to the four questions were coded independently, and the percentage of codes having a kappa ≥ 0.9 ranged from 38.5%–61.1% depending on the question ("Think back to when you discussed male condom use with your partner since the last session. What exactly did you ask/tell him?" [Q6A]: 38.5%; "How do you think your partner would react if you asked him to use male condoms?" [Q4A]: 50.0%; "Why can't you refuse sexual intercourse if your husband does not agree to use the male condom?" [Q7A]: 46.2%; "How did he react when you asked him to use the male condom?" [Q6C]: 61.1%).

Number of Rounds Required to Achieve Acceptable Intercoder Reliability

The hemophilia study data were coded by independent coders for only two rounds, and at the second round, only 64.6% (135/209) of coded themes had kappa ≥ 0.8. The LA bathhouse study required eight rounds to achieve 80% of codes having kappas ≥ 0.9. Within the Zimbabwe study, the four open-ended questions required different numbers of coding rounds to achieve most codes (90%) having a kappa greater than 0.9. Specifically, one question (Q6A) required nine rounds, while two questions required four rounds and one required three rounds (see Table 2). Even within this limited range of studies, we see wide variation in the number of rounds required to reach acceptable levels of intercoder reliability.

Factors Associated with Initial Intercoder Reliability and Fewer Coding Rounds

There was substantial variation in initial intercoder agreement and in the number of rounds required to achieve acceptable agreement. Discussions with coders coupled with the analysis of the coding process revealed factors that might influence these aspects of the process.

Number of codes per coding round. Coders observed that dealing with a large number of codes at any given coding round (e.g., approximately thirty with Zimbabwe study question 6a or more than 200 for the Hemophilia study) made coding decisions very difficult. It was therefore hypothesized that coding schemes with fewer codes would result in higher initial intercoder reliability and fewer rounds to achieve acceptable levels of intercoder reliability. To examine this possibility, the Zimbabwe study coding team restricted the number of possible codes (<20) when coding responses for three of the open-ended questions. Initial intercoder reliability improved (46.2%, 50.0%, and 61.1% of codes having kappa ≥0.9 for Zimbabwe study questions 7A, 4, 6C, respectively, compared to 38.5% ≥ 0.9 for Zimbabwe study question 6A and 32.9% ≥ 0.8 for the hemophilia study interview) when decreasing the number of codes (see Table 2). Furthermore, the number of rounds required to achieve acceptable levels of intercoder reliability was reduced from nine (for Zimbabwe study question 6a) to three–four rounds (for Zimbabwe study questions 7A, 4, 6C).

Length of text segments. Coders who worked with the LA bathhouse study data reported that the interviews' longer, unsegmented text units (up to one page of text) increased the complexity of coding decisions. Even though the LA bathhouse study generally used fewer than twenty codes for each coding decision, the initial intercoder reliability for the LA bathhouse study (39.0% of codes with kappa ≥ 0.9) was approximately as low as that for the hemophilia study and question 6a from the Zimbabwe study. The LA bathhouse data also required eight rounds to achieve less stringent levels of intercoder reliability (80% of codes having kappa ≥ 0.9).

Other factors. Coders from the various studies also noted other factors that seemed to compromise a team's ability to achieve high intercoder reliability. First, interview quality sometimes varied substantially by interviewer. Certain interviewers were more adept at keeping interview responses succinct and on task (related to study questions), and the texts transcribed from these interviews were considered easier to code. Second, coders from the LA bathhouse study noted that the initial codebook was generated prior to viewing any of the interviews. This initial codebook poorly reflected interview responses, and coders felt that several coding rounds could have been avoided if the initial codebook had been generated after first reading the set of texts.

DISCUSSION

The analyses in this article reveal two common patterns. First, two coders using the first draft of a codebook will generally show low agreement on individual codes. Second, through an iterative process of codebook revision and clarification, it is possible to remedy initially low agreement in a finite number of coding rounds. The analyses also provide tentative evidence that simple changes to procedures (reducing number of codes per round, reducing average text segment length per round, using question-specific codes, etc.) can increase initial intercoder reliability and reduce the number of coding rounds required to reach acceptable levels of intercoder reliability.

We believe that reliable coding of text, although not sufficient to guarantee the validity of conclusions drawn from text data, is a necessary criterion for ensuring quality control during the research process. The fact that two coders may differ greatly in their first coding of a text suggests that conclusions made by a lone interpreter of text may not reflect what others would conclude if allowed to examine the same set of texts. In other words, without checks from other interpreters, there is an increased risk of random error and bias in interpretation. This should be a cause for concern, but it does not mean that we should abandon hope of achieving good intercoder reliability. Indeed, we have shown that it is possible to generate a codebook that can be applied reliably by different coders. This is an important finding considering that applied and multidisciplinary research settings often require that researchers be able to vouch for the reliability of the process by which conclusions followed from their data.

The finding that initial codings of a text generate low levels of agreement between coders is consistent with findings from other studies (Carey 1996; Weinberger et al. 1998; Hagelin 1999) and perhaps explained by the numerous pathways by which coders can differ (Potter and Levine-Donnerstein 1999:271). It is also consistent with advice regarding what to expect during the iterative coding of text (Miles and Huberman 1994:64). Whereas some studies have stopped at the first round of coding (Weinberger et al. 1998), this article finds that a few coding iterations can generate acceptable agreement between coders, even when dealing with text data that has high interpretive burden (Willms et al. 1990; Carey, Morgan, and Oxtoby 1996; Hagelin 1999).

There are two possible explanations for why intercoder reliability improves with iterative between-coder comparisons and revisions to a codebook. First, the iterative process of codebook development successively clarifies terms and definitions so that (1) all members of a coding team can understand them and (2) redundant codes and codes with overlapping definitions are eliminated. Second, teams of coders may experience a kind of "interpretive convergence," where repeated discussions about the texts cause their interpretations to converge (Hak and Bernts 1996). The latter explanation implies that the procedures described in this articles do not actually generate a codebook for which all coders will have high intercoder reliability but rather create an interpretive framework that may only be specific to the current team of coders. The possibility of interpretive convergence arises most obviously in the LA bathhouse study, where all respondents were coded at each coding round. In this case, coders coded text segments that they had coded and discussed in previous rounds, and their consensus may have been the result of the discussion more than the result of a process that clarifies terms and code definitions. Consensus also can be influenced by supervision, interpersonal persuasion, conformity and other sources of "training effects," as well as adoption of a particular conceptual or policy framework to guide the analysis.

Future analyses will need to examine the degree to which "interpretive convergence" and training effects explain the increase in intercoder reliability observed in these studies. One method for cross-validating results would be to iteratively generate a final codebook for a set of texts using one set of coders and then having a second set of coders independently apply the codebook to the same set of texts, with supervision that is independent of the original coding. If the new coders agree with each other as much as the original coders did after their final round of coding, then we could argue that the iteratively revised codebook, rather than interpretive convergence, is what drives high intercoder reliability. If, on the other hand, the new coders do not agree any more than the original coders did on their initial round of coding, then we could argue that "reliable" coding generated by the original coders was a result of "interpretive convergence" specific to the original coders. High intercoder reliability likely would be the result of both of these processes.

Experience coding text from these three and several other studies suggests that steps can be taken to reduce the time required to achieve acceptable levels of intercoder reliability. These findings are consistent with observations by other researchers that large portions of text, such as paragraphs and complete texts, usually are more difficult to code as units then smaller portions, such as words and phrases, because large units typically contain more information and a greater diversity of topics (Weber 1990:16). They further coincide with findings in decision-making research that increasing the complexity of decision tasks (in this case with too many codes and too long text units) may decrease optimal decision making (Simon 1981).

By examining the intercoder reliability process across several studies, this article provides benchmarks about the flow of the process and suggestions about how to improve it. Several limitations of the available data, however, constrain the inferences we can make. We have considered the intercoder reliability process in only three studies, each conducted at different stages in the development of the reliability process. The conclusion about low initial reliabilities holds clearly throughout the studies, but further investigation is needed to confirm the role that (1) shorter text segment length and (2) smaller numbers of codes per coding round play in speeding up the generation of high intercoder reliability. We provide additional, more detailed guidelines for establishing intercoder reliability in the appendix at the end of this article.

Although we examined studies that vary with respect to response length, study purpose, and number of participants, each study did rely on short, semistructured text formats applied in a relatively uniform fashion to relatively large samples. These facts limit our ability to judge to what degree it could be extended to other forms of data, such as qualitative data generated from unstructured focus groups or open-ended ethnographic interviews (Morse 1997). At the same time, intercoder agreement has been achieved for coding of a wide range of data, including projective test data, photographs and other visual data, clinical chart reviews, and behavioral data on video (Weber 1990; Hagelin 1999; Bell 2001; Heyman et al. 2001). Moreover, other researchers at the CDC have been successful in achieving high levels of intercoder reliability in the analysis of unstructured qualitative data (MacQueen, McLelland, and Milstein 1998). One common theme throughout these lines of research is that a structured dialogue about a text (or other set of data) between at least two individuals is a good way of getting interpretations that do not depend solely on a lone observer. This tenet is based on the observation that individual decisions, even "expert decisions," are limited by a number of potential biases (i.e., the failure to use "disconfirmatory" evidence; Simon 1981) and random errors that can be partially corrected by well-organized discussions between individuals (Hutchins 1995). The discussion between coders described in this article is structured to reduce intercoder variance that may result either from random variation between coders or from systematic differences in the coders' interpretive frameworks (Murphy and De Shon 2000).

There are also several issues that we have not addressed in the process and that deserve further research. First, we have only considered studies where the coding team consisted of two members. Extensions of Cohen's kappa exist that allows for multiple coders, and further research should examine how the number of coders may effect the coding process (Fleiss 1971; Fleiss, Nee, and Landis 1979; Conger 1980; Fleiss 1981:225–34). Second, we have not assessed the "validity" of the codebooks generated by this process (Armstrong et al. 1997; Potter and Levine-Donnerstein 1999). We also have not examined whether another set of coders would arrive at similar codebooks if they followed the reliability process (intercoder reliability at the level of codebook rather than coding). Nor have we examined the degree to which the categories defined in each codebook would be recognizable to the populations we have studied. We have only begun to assess whether the variables created with this approach are associated with external criteria of interest to HIV research (contraceptive use outcomes, HIV incidence) (Hruschka et al. 2004). As mentioned earlier, we have not yet assessed to what degree interpretations generated by the reliability process are a product of (1) the convergence of coders' decision making ("interpretive convergence") versus (2) the increasing clarity of the codebook.

In this article, we are not advocating the generation of intercoder reliability as an end in itself or simply as a gesture to gain acceptability in the eyes of other disciplines. Rather, we view reliability as an important concept in many disciplines, not because it conveys a sense of respectability or prestige, but because it is a useful assessment of the quality of the process by which data become conclusions. Another consideration is that intercoder agreement is only one index of reliability (Cronbach, 1990; Cicchetti, 1994), and it does not address areas such as intracoder agreement (i.e., internal consistency) or the temporal reliability of the data being coded. Furthermore, we understand that the process proposed in this article may have limitations in its extension to other more complex forms of content. At the same time, if we want the information we produce from qualitative inquiry to remain useful and credible in applied and multidisciplinary settings, it will be important to develop means for assessing and guaranteeing the quality of that information. The process proposed in this article is one such attempt, and we hope that it serves either as a useful template for analysis or as an impetus to propose further improvements to quality assessment in the analysis of text.

APPENDIX
Guidelines

Based on experience coding qualitative data from these and other CDC-sponsored studies, we have further compiled a set of guidelines to improve the efficiency of coding and the generation of codebooks that facilitate high agreement between coders. These are intended as a supplement and revision to the eight suggestions already listed by MacQueen, McLelland, and Milstein (1998)—assign one lead coder; schedule regular coding team meetings; seek to enhance intercoder reliability; develop a plan for segmenting text while codebook is being developed; establish intercoder reliability measures early in coding process; in defining codes, do not assume anything is obvious; throw out codes that do not work and rework definitions of problematic codes; and accept the fact that text will need to be recoded.

1. *Generating good intercoder reliability will not fix poor study design or data quality*. Ensuring that coders have reliably coded a text is necessary but not sufficient for accurate and generalizable conclusions. Generating coding procedures that facilitate high intercoder reliability is only one link in a long research chain (including sampling, interviewer training, data collection, and statistical analysis), and poor design or preparation at any of these links can break the chain and compromise the accuracy or generalizability of one's final conclusions. As Morse (1997) has pointed out, if one has collected data in nonuniform ways or chosen a sample that can be said only to represent itself, establishing high intercoder reliability will purchase little for the researcher. Experience at the CDC suggests that interviewers who are not able to keep interviews on task generate data that are difficult to code reliably. Low quality data have little theoretical or applied value, regardless of the degree of rigor used during coding and analysis. It is therefore important to maintain the quality of all stages of the research process if assessments of intercoder reliability are to be feasible or useful.
2. *Clear physical organization of texts, codebooks, codes, and coders*. Coding is a difficult process, further complicated by the addition of multiple coders. A clear organization of texts, codebooks, and codes can help reduce this complexity. Computer packages (EZ-Text, AnSWR) can aid in the organization of texts, codebooks, and codes (Hohnloser, Kadlec, and Puerner 1995; Carey

et al. 1998, MacQueen, McLelland, and Milstein 1998; Strotman et al. 2002). Prior to generating the first codebook, the team should also organize itself by nominating one person (preferably someone with prior experience coding) who is in charge of (1) noting codebook revisions during team meetings, (2) making the revisions, and (3) distributing texts to coders during reliability assessments (MacQueen, McLelland, and Milstein 1998). This avoids the problem of having multiple versions of a codebook being used by different coders.

3. *At each coding round, code independent subsamples*. To reduce the possibility that intercoder reliability for a code is the result of previous discussions about specific text segments, coders should encounter previously uncoded text at each new coding round. This did not occur in the LA bathhouse study, which raises questions about the sources of intercoder agreement.
4. *Limit the number of codes*. MacQueen, McLelland, and Milstein (1998) suggest that the number of codes used in each iterative coding round be limited to between thirty and forty. Based on our experience, we suggest a further reduction to fewer than twenty codes, and coding only one question at a time. Our experience also indicates that global codes (codes that can be applied to any question) are more difficult for coders than question-specific codes. We therefore, advocate using global codes only when absolutely necessary.
5. *Do not sacrifice relevance or meaning for reliability*. As Rose and Webb (1998) point out, an attention to rigor should never stifle the creative element in data analysis. More importantly, researchers should take steps to ensure that in generating intercoder reliability, one does not sacrifice meaning. It is relatively easy to develop codes on which raters can consistently agree. At an extreme, one could construct a code ("THE") that is assigned whenever the word "the" appears in the text. Because of the simplicity of this recognition task, two coders would likely be able to generate very similar assignments of this code. The code "THE," therefore is easily coded in a reliable manner. It may be, however, of little use in understanding the substantive content of the data. While reliability is seen as a precondition for validity, it does not automatically confer evidence of validity (Altheide and Johnson 1994:487). To avoid the generation of codes that facilitate intercoder reliability but that mean nothing, it is important to constantly assess whether (1) each code is relevant to a research question and (2) the codebook definition of each code reflects what it is meant to capture.
6. *Do not assess intercoder reliability by solely using a simple percentage of agreement*. Although Cohen (1960) noted more than forty years ago that a simple percentage of agreement (# of code agreements/# of coding decisions) overestimates intercoder agreement by not taking chance agreement into account, numerous studies continue to use it as a measure of intercoder agreement (Kolbe and Burnett 1991; Lombard, Snyder-Duch, and Campanella 2002). A number of statistics, including the kappa statistic (Cohen 1960), have been proposed that do take chance agreement into account (see Banerjee

et al. 1999). Several concerns have been raised about the kappa statistic, including its dependence on code frequencies and its conservative estimation of intercoder agreement (Brennan and Prediger 1981; Byrt, Biship, and Carlin 1993; Googenmoos-Holzmann 1993). For ordinal or interval codes, there are a number of extensions and alternatives to the kappa (Banerjee et al. 1999). Nonetheless, sole use of the simple percentage of agreement is not an acceptable alternative.

7. *Choosing an appropriate cut-off for kappa*. Having chosen a reliability statistic, it then becomes necessary to determine appropriate levels of reliability. There is some variance in the guidelines that authors have constructed for evaluating kappa. Cicchetti (1994) and Fleiss (1981) have proposed the following criteria: 0.75–1.00 = *excellent*; 0.60–0.74 = *good*; 0.40–0.59 = *fair*; and < 0.40 is *poor*. Landis and Koch's (1977) earlier criteria were the following: 0.81–1.00 = *almost perfect*; 0.61–0.80 = *substantial*; 0.41–0.60 = *moderate*; 0.21–0.40 = *fair*; 0.00–0.20 = *slight*; and < 0.00 = *poor*. Miles and Hubermann (1994) do not specify a particular intercoder measure, but they do suggest that intercoder reliability should approach 0.90, with some consideration given to the range and size of coding schemes (Miles and Huberman 1994; Bernard 2002:483). Overall, there seems to be rough agreement on acceptable criteria for the highest levels of intercoder agreement, so we have proposed using kappas ≥ 0.80–0.90 as a target. It's important to consider that Cicchetti's criteria take into account clinical significance, a consideration similar to the policy and program applications that were being considered in the studies presented here. From our experience with coding short segments of text, it is possible to reach high levels of reliability for most codes (kappa ≥ 0.8 or kappa ≥ 0.9) using the iterative coding process. There are benefits and costs associated with the choice of a cutoff. If one is only concerned with a rough estimate of a code's population prevalence, a cutoff of kappa ≥ 0.8 or even kappa ≥ 0.7 could be acceptable. A disadvantage of higher cutoffs is the increased energy, time, and coding iterations that may be required to reach them. When the correct classification of individual cases is important, as in the case of clinical diagnosis or assignment to a particular treatment or intervention, stringent cutoffs (kappa ≥ 0.9) may be necessary.
8. *Clear presentation of results*. In a recent review of intercoder reliability procedures in communications research, Lombard, Snyder-Duch, and Campanella (2002) found little uniformity in presentation of results and sparse attention to issues of reliability of interpretations. Kolbe and Burnett (1991) found a similar inattention to reliability issues in a review of content analysis a decade earlier. In many cases, papers will mention intercoder reliability but then omit either exactly how they achieved intercoder reliability or what measure of reliability they achieved (Kolbe and Burnett 1991). Experience with analyzing qualitative data from the three studies discussed in this article has shown that it is possible to follow a series of clear, easily reportable steps to assess and establish intercoder reliability. Lombard, Snyder-Duch,

and Campanella (2002) provide clear guidelines for reporting on intercoder reliability. Other examples of options for reporting reliability statistics in tables can be found in Carey, Morgan, and Oxtoby (1996). Finally, when presenting data on individual codes, the code-specific kappa should be included because it may be very different from the summary measure used to assess the overall codebook. Those codes for which coders could not achieve high agreement should be interpreted cautiously.

REFERENCES

Altheide, D. L., and J. M. Johnson. 1994. Criteria for assessing interpretive validity in qualitative research. In *Handbook of qualitative research*, edited by N. K. Denzin and Y. Lincoln, 485–99. Thousand Oaks, CA: Sage.

Appleton, J. V. 1995. Analysing qualitative interview data: Addressing issues of validity and reliability. *Journal of Advanced Nursing* 22:993–7.

Armstrong, D., A. Gosling, J. Weinman, and T. Marteau. 1997. The place of inter-rater reliability in qualitative research: An empirical study. *Sociology* 31:597–606.

Banerjee, M., M. Capozzoli, L. McSweeney, and D. Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics* 27:3–23.

Bell, P. 2001. Semiotics and the content analysis of visual images. *Journal of Information, Communication and Library Sciences* 7:81–100.

Bernard, H. R. 2002. *Research methods in anthropology: Qualitative and quantitative approaches*. 3rd ed. Walnut Creek, CA: Altamira.

Boyatzis, R. E. 1998. *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

Brennan, R. L., and D. J. Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement* 41:687–99.

Byrt, T., J. Biship, and J. B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46 (5): 423–9.

Carey, J., P. H. Wenzel, C. Reilly, J. Sheridan, and J. M. Steinberg. 1998. CDC EZ-TEXT: Software for management and analysis of semistructured qualitative data sets. *Cultural Anthropology Methods Journal* 10:14–20.

Carey, J. W., M. Morgan, and M. Oxtoby. 1996. Inter-coder agreement in analysis of responses to open-ended interview questions: Examples From tuberculosis research. *Cultural Anthropology Methods* 8:1–5.

Cicchetti, D. V. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6:284–90.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychosocial Measurement* 20:37–46.

Conger, A. J. 1980. Integration and generalization of kappas for multiple raters. *Psychological Bulletin* 88:322–8.

Cronbach, L. J. 1990. *Essentials of psychological testing*. New York: Harper Collins.

Denzin, N. K., and Y. Lincoln. 1994. *Handbook of qualitative research*. Thousand Oaks, CA: Sage.

Drisko, J. 1997. Strengthening qualitative studies and reports: Standards to promote academic integrity. *Journal of Social Work Education* 33:185–97.

Field, A. P., and J. M. Morse. 1985. *Nursing research: The application of qualitative approaches.* London: Croom Helm.

Fitzpatrick, R., and M. Boulton. 1996. Qualitative research in health care I: The scope and validity of methods. *Journal of Evaluation in Clinical Practice* 2:123–30.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulleting* 76 (5): 378–82.

———. 1981. *Statistical methods for rates and proportions.* 2nd ed. New York: Wiley.

Fleiss, J. L., J. C. M. Nee, and J. R. Landis. 1979. Large sample variance of kappa in the cases of different sets of raters. *Psychological Bulletin* 86 (5): 974–7.

General Accounting Office (GAO). 1991. *Using structured interview techniques.* Washington DC: U.S. GAO.

Googenmoos-Holzmann, Irene. 1993. How reliable are chance-corrected measures of agreement? *Statistics in Medicine* 12:2191–205.

Gorden, R. 1992. *Basic interviewing skills.* Itasca, IL: F.E. Peacock.

Guba, E. G., and Y. S. Lincoln. 1994. Competing paradigms in qualitative research. In *Handbook of qualitative research*, edited by N. K. Denzin and Y. Lincoln, 105–17. Thousand Oaks, CA: Sage.

Hagelin, E. 1999. Coding data from child health records: The relationship between interrater agreement and interpretive burden. *Journal of Pediatric Nursing* 14:313–21.

Hak, T., and T. Bernts. 1996. Coder training: Theoretical training or practical socialization? *Qualitative Sociology* 19:235–57.

Heyman, R. E., B. R. Chaudhry, D. Treboux, J. Crowell, C. Lord, D. Vivian, and E. B. Waters. 2001. How much observational data is enough? An empirical test using marital interaction coding. *Behavior Therapy* 32:107–22.

Hohnloser, J. H., P. Kadlec, and F. Puerner. 1995. Experiments in coding clinical information: An analysis of clinicians using a computerized coding tool. *Computers and Biomedical Research* 28:393–401.

Hruschka, D., B. Cummings, D. Cobb St. John, J. Moore, G. Khumalo-Sakutukwa, and J. W. Carey. 2004. Fixed-choice and open-ended response formats: A comparison from HIV Prevention Research in Zimbabwe. *Field Methods* 16:184–202.

Hutchins, E. 1995. *Cognition in the wild.* Cambridge, MA. MIT Press.

Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Klahr, D., and K. Kotovsky. 1989. *Complex information processing: The impact of Herbert A. Simon*. Hillsdale, NJ: Lawrence Erlbaum.

Kolbe, R. H., and M. S. Burnett. 1991. Content-analysis research: An examination of applications with directions for improving research reliability and objectivity. *Journal of Consumer Research* 18:243–50.

Krippendorff, K. 1980. *Content analysis: An introduction to its methodology.* Beverly Hills, CA: Sage.

———. 1995. On the reliability of unitizing continuous data. *Sociological Methodology* 25:47–76.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–74.

Liebetrau, A. M. 1983. *Measures of association.* Newbury Park, CA: Sage.

Lombard, M., J. Snyder-Duch, and C. C. Campanella. 2002. Content analysis in mass communication assessment and reporting of inter-coder reliability. *Human Communication Research* 28 (4): 587–604.

MacQueen, K. M., E. McLelland, K. Kay, and B. Milstein. 1998. Codebook development for team-based qualitative analysis. *Cultural Anthropology Methods* 10:31–6.

Madill, A., A. Jordan, and C. Shirley. 2000. Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology* 91:1–20.

Mantell, J. E., A. T. DiVittis, and M. I. Auerbach. 1997. *Evaluating HIV prevention interventions*. New York: Plenum.

Mays, N., & Pope, C. 1995. Rigour and qualitative research. *British Medical Journal* 311:109–12.

———. 2000. Assessing quality in qualitative research. *British Medical Journal* 320:50–2.

Miles, M. B., and A. M. Huberman. 1994. *Qualitative data analysis*. Thousand Oaks, CA: Sage.

Morse, J. M. 1997. Perfectly healthy but dead: The myth of inter-rater reliability . *Qualitative Health Research* 7:445–7.

Murphy, K. R., R. De Shon. 2000. Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology* 53:873–900.

Mutchler, M. G., T. Bingham, M. Chion, R. A. Jenkins, L. E. Klosinski, and G. Secura. In press. *Comparing sexual behavioral patterns between two bathhouses: Implications for HIV prevention intervention policy. Journal of Homosexuality.*

Neumark-Sztainer, D., and M. Story. 1997. Recommendations from overweight youth regarding school-based weight control programs. *Journal of School Health* 67:428–33.

Nunnally, J. C., and I. H. Bernstein. 1994. *Psychometric theory*. New York: McGraw-Hill.

O'Leary, A., J. Moore, G. Sakatukwa, L. Loeb, D. Hruschka, R. Khan, and N. Padian. 2003. Association of negotiation strategies to consistent use of male condoms by women receiving an HIV prevention intervention in Zimbabwe. *AIDS* 17:1705-7.

Parsons, J. T., H. C. Huszti, S. O. Crudder, B. Gage, D. Jarvis, J. Mendoza, and K. L. Parish. 1998. Determinants of HIV risk reduction behaviors among female partners of men with hemophilia and HIV infection. *AIDS and Behavior* 2:1–11.

Potter, W.J., and D. Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research* 27:258–84.

Rose, K., and C. Webb. 1998. Analyzing data: Maintaining rigor in a qualitative study. *Qualitative Health Research* 8:556–62.

Ryan, G. W., and H. R. Bernard. (2003). Techniques to identify themes. *Field Methods* 15 (1): 85–109.

Simon, H. A. 1981. *The sciences of the artificial*. 2nd ed. Cambridge, MA: MIT Press.

Strotman, R., E. McLelland, K. M. MacQueen, and B. Milstein. 2002. AnSWR: Analysis software for Word-based records, version 6.2. Atlanta: Centers for Disease Control and Prevention.

Sykes, W. 1991. Taking stock: Issues from the literature on validity and reliability in qualitative research. *Journal of the Market Research Society* 33:3–12.

Wang, S. S. L., L.-C.Lin, and B. Ing-Tau Kuo. 1997. The health care needs of hospitalized patients with AIDS in Taiwan. *AIDS Patient Care and STDS* 11:179–88.

Weber, R. P. 1990. *Basic content analysis*. Newbury Park, CA: Sage.

Weinberger, M., J. A. Ferguson, G. Westmoreland, L. A. Mamlin, D. S. Segar, G. J. Eckert, J. Y. Greene, D. K. Martin, and W. Tierney. 1998. Can raters consistently evaluate the content of focus groups? *Social Science and Medicine* 46:929–33.

Willms, D. G., J. A. Best, D. W. Taylor, J. R. Gilbert, D. M. C. Wilson, E. A. Lindsay, and J. Singer. 1990. A systematic approach for using qualitative methods in primary prevention research. *Medical Anthropology Quarterly* 4:391–409.

*DANIEL J. HRUSCHKA, MPH, is a doctoral candidate in the Department of Anthropology at Emory University, Atlanta, Georgia. His dissertation research focuses on the development of social networks and activity preferences during middle school in the United States. Recent publications include (with B. Cummings, D. Cobb St. John, J. Moore, G. Khumalo-Sakutukwa, and J. Carey) “Fixed-Choice and Open-Ended Response Formats: A Comparison from HIV-Prevention Research in Zimbabwe” (*Field Methods*, 2004).*

DEBORAH SCHWARTZ, MA, is a behavioral scientist in the Prevention Research Branch of the Division of HIV/AIDS Prevention at the U.S. Centers for Disease Control and Prevention in Atlanta, Georgia. She currently helps lead a multisite research study exploring contextual factors associated with recent HIV infection. She is coauthor of the “HIV/AIDS Research and Prevention: Anthropological Contributions and Future Directions” chapter (with J. Carey, M. Spink-Neumann, D. Easton, E. Picone-DeCaro, and D. Cobb St. John) in the Encyclopedia of Medical Anthropology *(2003).*

DAPHNE COBB ST. JOHN, MPH, is a public health advisor with the Epidemiology Branch, Division of HIV/AIDS, Centers for Disease Control and Prevention (CDC), in Atlanta, Georgia. Since 1995, she has worked on several multidisciplinary intervention research and evaluation projects that focused on topics such as HIV prevention, contraceptive use, and women’s health promotion. She currently supports international microbicide and vaccine trial efforts in Botswana and Thailand. She is coauthor of the HIV and anthropology chapter in the Encyclopedia of Medical Anthropology *(2003).*

ERIN PICONE-DECARO, MPH, is a health scientist with the Program Evaluation Research Branch, Division of HIV/AIDS, Centers for Disease Control and Prevention (CDC). Since joining CDC in 2000, she has worked on multidisciplinary studies of the context of recent HIV infection, qualitative and quantitative evaluation of behavior change interventions, and systematic review of HIV behavioral interventions. Prior work included a prospective cohort study of infant feeding behaviors among new mothers in New York. Recently, she has been coauthor of the HIV and anthropology chapter in the Encyclopedia of Medical Anthropology *(2003).*

RICHARD A. JENKINS, PhD, is a behavioral scientist with the Prevention Research Branch, Division of HIV/AIDS Prevention, Centers for Disease Control and Prevention (CDC), in Atlanta, Georgia. His research has focused on the social and behavioral epidemiology of HIV in Thailand, preparations for HIV vaccine trials, and development of ways to increase data use by community planning groups. His current activities include projects looking at factors related to seronversion among men who have sex with men and application of novel data collection methods including computer-assisted interviews and rapid HIV tests with varied populations in Thailand.

JAMES W. CAREY, PhD, MPH, is the section chief for the Methods Research Section within the Prevention Research Branch (part of the Division of HIV/AIDS Prevention at the Centers for Disease Control and Prevention [CDC], Atlanta, Georgia). Since joining the CDC in 1992, he has directed numerous multidisciplinary research projects aimed at understanding individual- and community-level factors influencing HIV risk behaviors, tuberculosis control, and effective public health program design in the United States and in other countries. A selection of his recent publications includes “Sexual Health Risks among Young Thai Women: Implications for HIV/STD Prevention and Contraception” (with D. Allen, R. Jenkins, et al. in AIDS and Behavior, *2003); “Condom Use among Vocational School Students in Chiang Rai, Thailand” (with R. Jenkins, C. Manopaiboon, et al. in* AIDS Education and Prevention, *2002); “HIV Prevention in Primary Care: Impact of a Clinical Intervention” (with J. Blue Spruce, W. Dodge, et al. in* AIDS Patient Care and STDs, *2001).*