How Many Interviews Are Enough? An Experiment with Data Saturation and Variability

GREG GUEST
ARWEN BUNCE
LAURA JOHNSON
Family Health International

Guidelines for determining nonprobabilistic sample sizes are virtually nonexistent. Purposive samples are the most commonly used form of nonprobabilistic sampling, and their size typically relies on the concept of "saturation," or the point at which no new information or themes are observed in the data. Although the idea of saturation is helpful at the conceptual level, it provides little practical guidance for estimating sample sizes, prior to data collection, necessary for conducting quality research. Using data from a study involving sixty in-depth interviews with women in two West African countries, the authors systematically document the degree of data saturation and variability over the course of thematic analysis. They operationalize saturation and make evidence-based recommendations regarding nonprobabilistic sample sizes for interviews. Based on the data set, they found that saturation occurred within the first twelve interviews, although basic elements for metathemes were present as early as six interviews. Variability within the data followed similar patterns.

Keywords: *interviewing; saturation; variability; nonprobability sampling; sample size; purposive*

While conducting a literature review of guidelines for qualitative research in the health sciences, we were struck by how often the term *theoretical saturation* arose. Article after article recommended that purposive sample sizes be determined by this milestone (e.g., Morse 1995; Sandelowski 1995; Bluff 1997; Byrne 2001; Fossey et al. 2002), and a good number of journals in the

---

Financial support for this research was provided by the U.S. Agency for International Development through Family Health International, although the views expressed in this article do not necessarily reflect those of either organization. The authors thank Kerry McLoughlin (Family Health International), Betty Akumatey (University of Ghana, Legon), and Lawrence Adeokun, (Association for Reproductive and Family Health, Ibadan, Nigeria). Without their hard work, this article would not have been possible.

Field Methods, Vol. 18, No. 1, February 2006 59–82
DOI: 10.1177/1525822X05279903

health sciences require that theoretical saturation be a criterion by which to justify adequate sample sizes in qualitative inquiry. Saturation has, in fact, become the gold standard by which purposive sample sizes are determined in health science research.

Equally striking in our review was that the same literature did a poor job of operationalizing the concept of saturation, providing no description of how saturation might be determined and no practical guidelines for estimating sample sizes for purposively sampled interviews. This dearth led us to carry out another search through the social and behavioral science literature to see if, in fact, any generalizable recommendations exist regarding nonprobabilistic sample sizes. After reviewing twenty-four research methods books and seven databases, our suspicions were confirmed; very little headway has been made in this regard. Morse's (1995:147) comments succinctly sum up the situation; she observed that "saturation is the key to excellent qualitative work," but at the same time noted that "there are no published guidelines or tests of adequacy for estimating the sample size required to reach saturation."

Our experience, however, tells us that it is precisely a general, numerical guideline that is most needed, particularly in the applied research sector. Individuals designing research—lay and experts alike—need to know how many interviews they should budget for and write into their protocol, before they enter the field. This article is in response to this need, and we hope it provides an evidence-based foundation on which subsequent researchers can expand. Using data from a study involving sixty in-depth interviews with women in two West African countries, we systematically document the degree of data saturation and variability over the course of our analysis and make evidence-based recommendations regarding nonprobabilistic sample sizes.

We stress here that we intentionally do not discuss the substantive findings from our research; they will be presented elsewhere. This is a methodological article, and we felt that including a discussion of our study findings would be more distracting than informative. We do provide some background for the study, but our focus is mainly on the development and structure of our codebook and its evolution across the analysis process.

NONPROBABILISTIC AND PURPOSIVE SAMPLING

Calculating the adequacy of probabilistic sample sizes is generally straightforward and can be estimated mathematically based on preselected parameters and objectives (i.e., *x* statistical power with *y* confidence intervals). In

theory, all research can (and should when possible) use probabilistic sampling methodology, but in practice, it is virtually impossible to do so in the field (Bernard 1995:94; Trotter and Schensul 1998:703). This is especially true for hard-to-reach, stigmatized, or hidden populations.

Research that is field oriented in nature and not concerned with statistical generalizability often uses nonprobabilistic samples. The most commonly used samples, particularly in applied research, are purposive (Miles and Huberman 1994:27). Purposive samples can be of different varieties—Patton (2002), for example, outlined sixteen types of purposive samples—but the common element is that participants are selected according to predetermined criteria relevant to a particular research objective. The majority of articles and books we reviewed recommended that the size of purposive samples be established inductively and sampling continue until "theoretical saturation" (often vaguely defined) occurs. The problem with this approach, however, is that guidelines for research proposals and protocols often require stating up front the number of participants to be involved in a study (Cheek 2000). Waiting to reach saturation in the field is generally not an option. Applied researchers are often stuck with carrying out the number of interviews they prescribe in a proposal, for better or worse.[1] A general yardstick is needed, therefore, to estimate the point at which saturation is likely to occur.

Although numerous works we reviewed explain how to select participants (e.g., Johnson 1990; Trotter 1991) or provide readers with factors to consider when determining nonprobabilistic sample sizes (Miles and Huberman 1994; Bernard 1995; Morse 1995; Rubin and Rubin 1995; Flick 1998; LeCompte and Schensul 1999; Schensul, Schensul, and LeCompte 1999; Patton 2002), we found only seven sources that provided guidelines for actual sample sizes. Bernard (2000:178) observed that most ethnographic studies are based on thirty-sixty interviews, while Bertaux (1981) argued that fifteen is the smallest acceptable sample size in qualitative research. Morse (1994:225) outlined more detailed guidelines. She recommended at least six participants for phenomenological studies; approximately thirty-fifty participants for ethnographies, grounded theory studies, and ethnoscience studies; and one hundred to two hundred units of the item being studied in qualitative ethology. Creswell's (1998) ranges are a little different. He recommended between five and twenty-five interviews for a phenomenological study and twenty-thirty for a grounded theory study. Kuzel (1992:41) tied his recommendations to sample heterogeneity and research objectives, recommending six to eight interviews for a homogeneous sample and twelve to twenty data sources "when looking for disconfirming evidence or trying to achieve maximum variation." None of these works present evidence for their recommen-

dations. The remaining two references—Romney, Batchelder, and Weller (1986) and Graves (2002)—do provide rationale for their recommendations for quantitative data and are discussed later in the article.

STUDY BACKGROUND

The original study for which our data were collected examined perceptions of social desirability bias (SDB) and accuracy of self-reported behavior in the context of reproductive health research. Self-reports are the most commonly used measure of sexual behavior in the health sciences, and yet concern has been raised about the accuracy of these measures (Brody 1995; Zenilman et al. 1995; Weinhardt et al. 1998; Schwarz 1999; Weir et al. 1999; Crosby et al. 2002). One of the key factors identified as affecting report accuracy is a participant's concern with providing socially desirable answers (Paulhus 1991; Geary et al. 2003). Our study was therefore designed to inform HIV research and intervention programs that rely on self-reported measures of sexual behavior.

Using semistructured, open-ended interviews, we examined how women talk about sex and their perceptions of self-report accuracy in two West African countries—Nigeria and Ghana. We solicited suggestions for reducing SDB and improving self-report accuracy within various contexts. In addition, we asked participants to provide feedback regarding methods currently used to mitigate SDB within the context of HIV research and prevention, such as audio-computer-assisted self-interviews (ACASI) and manipulating various aspects of the interview.

METHODS

Sampling and Study Population

A nonprobabilistic, purposive sampling approach was used. We wanted to interview participants at high risk for acquisition of HIV and who would be appropriate candidates for HIV prevention programs in the two study sites. We therefore interviewed women who met at least three basic criteria: (1) were eighteen years of age or older, (2) had vaginal sex with more than one male partner in the past three months, and (3) had vaginal sex three or more times in an average week. Women at the highest risk for HIV in Nigeria and Ghana tend to be engaged in some form of sex work (although not all self-identify as sex workers), so fieldworkers recruited sex workers for our study.

TABLE 1
Sample Characteristics

	Ghana (n = *30*)	*Nigeria* (n = *30*)	*Combined* (N = *60*)
Age			
Mean	26.3	32.0	29.1
Range	20–35	1–53	19–53
Years of education			
Mean	6.8	10.3	8.5
Range	0–12	0–17	0–17
Number of ethnic groups	12	3	15
Marital status			
Single	20 (66.7%)	13 (43.3%)	33 (55%)
Married	0	1 (3.3%)	1 (1.7%)
Divorced/separated	10 (33.3%)	14 (46.7%)	24 (40%)
Widowed	0	2 (6.7%)	2 (3.3%)
Previous research experience	13 (43.3%)	6 (20%)	19 (31.7%)

In Nigeria, thirty high-risk women were recruited from three sites in the city of Ibadan, which correspond to different socioeconomic environments: brothels, a college campus, and known pick-up points for sex workers. The sampling process was similar in Ghana. Thirty high-risk women were recruited from greater Accra. Three high-risk sites were identified for recruitment and included a red light area, a hotel, and a hostel. Table 1 presents the sample characteristics for the two sites.

Data Collection and Analysis

The interview guide consisted of six structured demographically oriented questions, sixteen open-ended main questions, and fourteen open-ended subquestions. Subquestions were asked only if a participant's response to the initial question did not cover certain topics of interest. All respondents were asked identical questions in the same sequence, but interviewers probed inductively on key responses. The guide was divided into the following six domains of inquiry:

- perceptions of sexually oriented research,
- discussion of sex and condoms within the community (i.e., among peers),
- discussion of sex and condoms within the research context,
- interviewer characteristics,
- remote interviewing techniques (ACASI, phone interviews), and
- manipulating the environment of an interview.

Data were collected between September 15 and December 12, 2003. Interviews were conducted in English, Twi, and Ga in Ghana and in English, Pidgin English, and Yoruba in Nigeria. All interviews were tape recorded, and verbatim responses to each question were translated and transcribed by local researchers, using a standardized transcription protocol (McLellan, MacQueen, and Niedig 2003). Transcripts were reviewed by the principal investigator at each site for translation accuracy and revised when necessary. Thematic analysis was performed on the translated transcripts using Analysis Software for Word-based Records (AnSWR; Centers for Disease Control and Prevention 2004).

A codebook was developed by two data analysts, using a standard iterative process (MacQueen et al. 1998). In this process, each code definition has five parts: (1) a "brief definition" to jog the analyst's memory; (2) a "full definition" that more fully explains the code; (3) a "when to use" section that gives specific instances, usually based on the data, in which the code should be applied; (4) a "when not to use" section that gives instances in which the code might be considered but should not be applied (often because another code would be more appropriate); and (5) an "example" section of quotes pulled from the data that are good examples of the code.

In our analysis, the lead analyst created an initial content-based coding scheme for each set of six interviews. Intercoder agreement was assessed for every third interview using combined segment-based Kappa scores run on two double-coded transcripts (Carey, Morgan, and Oxtoby 1996). Coding discrepancies (individual codes receiving Kappa scores of 0.5 or less) were discussed and resolved by the analysis team, the codebook revised accordingly, and recoding performed when necessary to ensure consistent application of codes. The resulting overall Kappa score, by individual question, was 0.82. To identify key themes, we ran frequency reports in AnSWR.

THE EXPERIMENT

The Methods section refers to the procedures we used in our substantive analysis, yet these procedures did not provide us with the data we needed to determine thematic development and evolution over time and eventually the point at which saturation occurred in our data. We had to develop additional procedures and methods to operationalize and document data saturation.

Saturation can be of various types, with the most commonly written about form being "theoretical saturation." Glaser and Strauss (1967:65) first defined this milestone as the point at which "no additional data are being found whereby the (researcher) can develop properties of the category. As he

sees similar instances over and over again, the researcher becomes empirically confident that a category is saturated . . . when one category is saturated, nothing remains but to go on to new groups for data on other categories, and attempt to saturate these categories also."

For serious practitioners of the grounded theory approach, the term *theoretical saturation* refers specifically to the development of theory. Theoretical saturation occurs when all of the main variations of the phenomenon have been identified and incorporated into the emerging theory. In this approach, the researcher deliberately searches for extreme variations of each concept in the theory to exhaustion.

Although *theoretical saturation* is the most commonly used term in published works, frequency of use within multiple bodies of literature has resulted in its meaning becoming diffuse and vague. To avoid propagating this transgression, we rely on a more general notion of data saturation and operationalize the concept as the point in data collection and analysis when new information produces little or no change to the codebook. We wanted to find out how many interviews were needed to get a reliable sense of thematic exhaustion and variability within our data set. Did six interviews, for example, render as much useful information as twelve, eighteen, twenty-four, or thirty interviews? Did any new themes, for example, emerge from the data gathered between interview thirteen and interview thirty? Did adding thirty more interviews from another country make any difference?

To answer these questions, we documented the progression of theme identification—that is, the codebook structure—after each set of six interviews, for a total of ten analysis rounds.[2] We monitored the code network and noted any newly created codes and changes to existing code definitions. We also documented frequency of code application after each set of six interviews was added. The reasoning behind this latter measure was to see if the relative prevalence of thematic expression across participants changed over time. It could be the case, to take a hypothetical example, that one code in the first round of analysis was applied to all six of the transcripts from one site, implying an initial high prevalence across participants. It could also be true that the same code was never applied in the remaining twenty-four transcripts and that another code emerged for the first time in the seventh transcript and was applied to the rest of the transcripts for a frequency of twenty-four. We needed to assess this variability.

We created a cumulative audit trail, updating our records after analysis of each set of six transcripts. So, in our first analysis, we analyzed the first six transcripts, then added six more in our second analysis for an n of twelve, and so on. We started with the data from Ghana and kept adding six transcripts until we had completed all thirty interviews from this site. Six transcripts

from Nigeria were then added to the analysis for an *n* of 36, and the process was repeated until all sixty of the interviews from both sites had been analyzed and the code definitions finalized. In all, we completed ten successive and cumulative rounds of analysis on sets of six interviews. Internal codebook structure (conceptually relating codes to one another) was not manipulated until all of the base codes had been identified and all sixty transcripts coded.

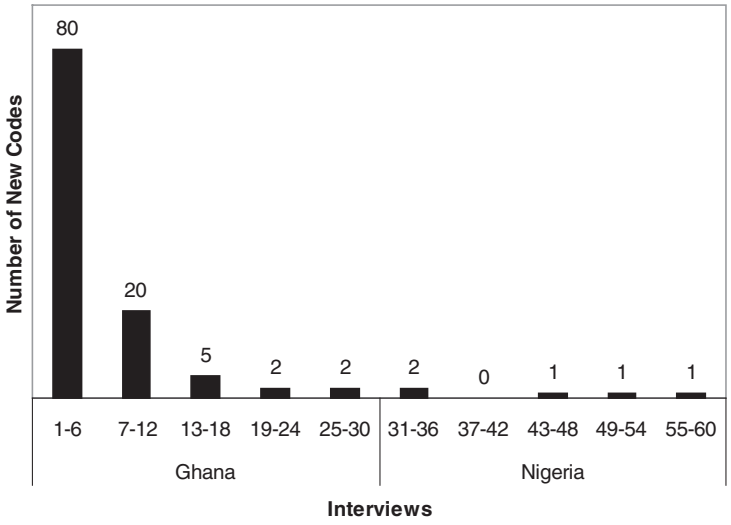
Below, we provide a summary of these data. Specifically, we present data illustrating the point in analysis when codes were created or definitions changed. We also examine frequency of code application across participants and estimate the point at which the distribution of code frequency stabilized. For all of our analyses, the unit of analysis is the individual participant (i.e., transcript) and data items the individual codes (i.e., expressions of themes).

Code Development

After analyzing all thirty interviews from Ghana, the codebook contained a total of 109 content-driven codes, all of which had been applied to at least one transcript. Of these codes, 80 (73%) were identified within the first six transcripts. An additional 20 codes were identified in the next six transcripts, for a cumulative total of 100, or 92% of all codes applied to the Ghana data. As one would expect, the remaining 9 codes were identified with progressively less frequency (see Figure 1, the five columns on the left, interviews 1–30). Clearly, the full range of thematic discovery occurred almost completely within the first twelve interviews—at least based on the codebook we developed (more on this later).

Surprisingly, not much happened to the number of codes once we started adding data from the other country. Only five new codes (out of a total of 114) had to be created to accommodate the Nigerian data (see Figure 1, the five columns on the right, interviews 31–60), one of which was new in substance. Four of the five codes were nonsubstantive in nature but were created to capture nuances in the Nigerian data. Two of these four new codes were needed for the unique subgroup of campus-based sex workers in Nigeria who typically do not refer to themselves or their friends as sex workers or to their sexual partners as clients. We therefore needed new codes that were the equivalent of codes used in other transcripts for talk among sex worker friends and talk about sex worker clients, but without the reference to sex workers. The result was two codes that were not new in substance but rather variations of the original codes tailored to the specific situation of the campus-based women. The two other codes were related to researcher qualities. One covered nonspecific statements that researchers' behavior and attitudes

FIGURE 1
Code Creation over the Course of Data Analysis



were important. The other code was akin to an "other" category and captured infrequently mentioned interviewer qualities.

Code Definition Changes

Creating a team-based codebook undoubtedly requires making changes to code definitions as data analysis progresses (MacQueen et al. 1998). Since this process itself will ultimately affect the absolute range of thematic expression identified in the data (e.g., if a code is augmented to be more inclusive of certain concepts), we documented all code definition changes throughout the analysis.

Table 2 illustrates the progression of definition changes throughout the analysis process. A total of thirty-six code revisions occurred throughout the entire analysis. The majority of changes (seventeen) occurred during the second round of analysis, and after analyzing only twelve interviews, 58% of the total number of changes had been made. Twenty-two of the thirty-six changes were singular in nature; definitions were revised twice for seven of the codes. No definitions were revised more than twice throughout the entire analysis.

TABLE 2
Code Definition Changes by Round of Analysis

	Analysis Round	*Interviews Analyzed*	*Definition Changes in Round*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
Ghana data						
	1	6	4	11	4	11
	2	12	17	47	21	58
	3	18	7	20	28	78
	4	24	3	8	31	86
	5	30	2	6	33	92
	6	36	3	8	36	100
Nigeria data						
	7	42	0	0	36	100
	8	48	0	0	36	100
	9	54	0	0	36	100
	10	60	0	0	36	100

As with the code creation data, adding the Nigerian data to the mix rendered little change to the codebook structure, with only three definition changes across thirty interviews, and these three changes occurred within the first six transcripts from this country. It appears that by the time we began looking at the Nigeria data, the structure of the codebook had been relatively well established, and incoming data offered few novel insights.

Codebook definitions also did not change much from a qualitative perspective. Of the thirty-six total code revisions, twenty-eight (78%) were made only to the "when to use" section of the definition, indicating that the substance of the code definition itself did not really change. Rather, clarifications and instructions for application were made more explicit. Of the ten codes whose actual definition changed over the course of the project, seven of the changes made the code more inclusive, thus expanding the conceptual scope of the definition. For example, the code "religion," which refers to "statements of a religious imperative to tell the truth," was changed to include both the positive and negative effects of religion on the veracity of self-reported behavior after analyzing the first set of transcripts from Nigeria (see Table 3). Three of the ten definition revisions narrowed the scope of the definition.

In Table 3, we present the definitions and subsequent revisions for the seven codes that were revised twice to provide examples of how codes were changed. Space constraints prohibit listing all code changes. Italics indicate the changes in definition. The number in parentheses at the end of each full

(continued on page 72)

TABLE 3
Sample Code Definition Revisions

	Original	*Revision 1*	*Revision 2*
Trust necessary (52%)	*Full definition:* Discussion of the necessity of trust before participant will discuss sexual matters with another woman; if there is no trust, will not talk about it *When to use:* Have to know somebody's "character" before talking personally; can't trust just anybody and often don't know most of the women well (R1)	*Full definition:* Discussion of the necessity of trust before participant will discuss sexual matters with another woman; if there is no trust, will not talk about it *When to use:* Have to know somebody's "character" before talking personally; can't trust just anybody and often don't know most of the women well; includes statements on confidences coming back to haunt you (R2)	*Full definition:* Discussion of the necessity of trust before discussing sexual matters with another woman; if there is no trust, will not talk about it *When to use:* Have to know somebody's "character" before talking personally; can't trust just anybody and often don't know most of the women well; includes statements on confidences coming back to haunt you; includes setting the right atmosphere (sharing own personal information, creating a safe space, etc.) before asking personal questions (R6)
Talk with sex worker (SW) friends (73%)	*Full definition:* Talk between the participant and a close friend who is a SW *When to use:* Often occurs as: can only share very personal matters with one to two close friends (concept of space, can go to their rooms but nobody else's) (R1)	*Full definition*: Talk between the participant and a close friend who is a SW *When to use:* If using the word *friend* in this discussion, use this code; often occurs as: can only share very personal matters with one to two close friends (concept of space, can go to their rooms but nobody else's) (R1)	*Full definition:* Talk between the participant and a close friend who is a SW *When to use:* If using the word *friend* in this discussion, use this code; often occurs as: can only share very personal matters with one to two close friends (concept of space, can go to their rooms but nobody else's); use if obvious/implied that the friend is a SW, even if not specifically stated (R2)

(continued)

TABLE 3 (continued)

	Original	*Revision 1*	*Revision 2*
Talk with non-SWs (15%)	*Full definition:* Talk among the participants and people who are not SWs (R2)	*Full definition:* Talk among the participants and people who are not SWs *When to use:* Includes both participant stating that she cannot talk to non-SWs about sex and/or condoms and statements that non-SWs don't talk about sex and/or condoms (R2)	*Full definition:* Talk among the participants and people who are not SWs *When to use:* Includes both participant stating that she cannot talk to non-SWs about sex and/or condoms and statements that non-SWs don't talk about sex and/or condoms *When not to use:* DO NOT use for statements that can talk about sexual issue with a non-SW (that would be coded as "cwho_sep") (R3)
Religion (25%)	*Full definition:* Statements of a religious imperative to tell the truth (even if not linked to gain) (R1)	*Full definition:* Statements of a religious imperative to tell the truth (even if not linked to gain) *When to use:* Includes people not being honest because they have turned away from God or are "wicked" (if this is meant in a religious sense) (R3)	*Full definition:* Religion or religious conviction and beliefs affect a person's honesty when discussing sexual issues *When to use:* Includes statements of a religious imperative to tell the truth (even if not linked to gain); includes people not being honest because they have turned away from God or are "wicked" (if this is meant in a religious sense); includes religion as a barrier to being honest about sexual issues (R6)
Personal help (62%)	*Full definition:* Being honest while in the study to receive personal help (advice, getting out) *When to use:* Can be moral, religious, and/or pragmatic, often all in same statement (R1)	*Full definition:* Being honest while in the study in order to receive personal help (advice, getting out) *When to use:* Can be moral, religious, and/or pragmatic, often all in same statement; includes statements that will be honest because the answers might help someone else (R2)	*Full definition:* Being honest while in the study to receive personal help (advice, getting out) *When to use:* Can be moral, religious, and/or pragmatic, often all in same statement; includes statements that will be honest because the answers might help someone else; the "help" includes learning (R3)

Reputation (55%)	*Full definition:* Reputation and self-image (don't want to be shamed/embarrassed/laughed at, do want to seem popular) drives respondent's (dis)honesty *When to use:* Sometimes give higher numbers because they want to seem "pretty" and popular; sometimes give lower numbers because they are embarrassed by how many men they sleep with in a day and don't want to seem greedy or money hungry; also code here at the top level if the respondent says, "I'll tell the truth because otherwise I'll be caught out in my lies" (R1)	*Full definition:* Reputation and self-image (don't want to be shamed/embarrassed/laughed at, do want to seem popular) drives respondent's (dis)honesty *When to use:* Sometimes give higher numbers because they want to seem "pretty" and popular; sometimes give lower numbers because they are embarrassed by how many men they sleep with in a day and don't want to seem greedy or money hungry; also code here at the top level if the respondent says, "I'll tell the truth because otherwise I'll be caught out in my lies"; cannot be honest about not using condoms because they know that it is "proper" to use them (R2)	*Full definition:* Reputation and self-image (don't want to be shamed/embarrassed/laughed at, do want to seem popular) drives respondent's (dis)honesty *When to use:* Sometimes give higher numbers because they want to seem "pretty" and popular; sometimes give lower numbers because they are embarrassed by how many men they sleep with in a day and don't want to seem greedy or money hungry or because they are "shy"; also code here at the top level if the respondent says, "I'll tell the truth because otherwise I'll be caught out in my lies"; cannot be honest about not using condoms because they know that it is "proper" to use them (R3)
Sex work stigma (17%)	*Full definition:* Statements on the stigma of being a SW *When to use:* Reactions from others ("not a human being"; Ghanaians very insulting about it, accused of bringing AIDS to Ghana) (R1)	*Full definition:* Statements on the stigma of being a SW *When to use:* Reactions from others ("not a human being"; Ghanaians very insulting about it, accused of bringing AIDS to Ghana); if talk about SW job as "disgraceful" (implies public stigma) (R1)	*Full definition:* Statements on the stigma of being a SW *When to use:* Reactions from others ("not a human being"; Ghanaians very insulting about it, accused of bringing AIDS to Ghana); if talk about SW job as "disgraceful" (implies public stigma); includes perceived stigma (e.g., talk about hiding while doing this job) (R2)

definition version indicates the round of analysis in which the code was originally created (in the Original column) and when the revision was made (in the Revision columns). Percentages in parentheses under the code name indicate the proportion of the sixty transcripts to which the code was applied.

Thematic Prevalence

Another critical dimension we needed to address was the overall relative importance of codes, for if codes developed in the early stages turned out to be the most important, doing additional interviews would tend to seriously diminish returns on time (and money) invested in additional interviews. Here, we define the importance of a code as the proportion of individual interviews to which a code is applied. We make the assumption that the number of individuals independently expressing the same idea is a better indicator of thematic importance than the absolute number of times a theme is expressed and coded. After all, one talkative participant could express the same idea in twenty of her responses and increase the overall absolute frequency of a code application significantly.

The first question we asked with respect to code frequency was at what point did relative frequency of code application stabilize, if at all? To assess this, we used Cronbach's alpha to measure the reliability of code frequency distribution as the analysis progressed. We present alpha values between each successive round of analysis, with each round containing an additional set of six interviews (see Table 4). The data transition point from one country to the next is also noted. For the Cronbach's alpha, .70 or higher is generally considered an acceptable degree of internal consistency (Nunnally and Bernstein 1994).

The data in Table 4 show that the alpha value is above .70 between the first two sets of interviews and that reliability of code frequency distribution increases as the analysis progresses, with the greatest increase occurring when the third group of interviews (interviews 13–18) are added. The consistency of application frequency appears to hold even when adding the interviews from Nigeria. In fact, internal consistency is higher for all ten rounds (i.e., both sites) combined (.9260) than for either of the five rounds of data analysis exclusive to each site (Ghana = .8766, Nigeria = .9173). Also, when we average code frequencies for each site and compare the two distributions, reliability between the two data sets remains high with an alpha of .8267.

Another question we had concerned the frequency dynamics associated with high prevalence codes, that is, codes applied to many transcripts. Did,

TABLE 4
Internal Consistency of Code Frequencies

	Rounds	*Interviews*	*Cronbach's Alpha*
Ghana only			
	1–2	1–12	.7048
	1–3	1–18	.7906
	1–4	1–24	.8458
	1–5	1–30	.8766
Ghana and Nigeria			
	1–6	1–36	.8774
	1–7	1–42	.8935
	1–8	1–48	.9018
	1–9	1–54	.9137
	1–10	1–60	.9260
μGhana, μNigeria			
	1–30, 31–60	.8267	

for example, themes that appeared to be important after six or twelve interviews remain important after analyzing all sixty interviews? Using the categorize function in SPSS, we transformed code frequencies into three groups: low, medium, and high. Based on these data, we found that the majority of codes that were important in the early stages of analysis remained so throughout. Of the twenty codes that were applied with a high frequency in round 1 of the analysis, fifteen (75%) remained in this category throughout the analysis. Similarly, twenty-six of the thirty-one high-frequency codes (84%) in the second round of analysis (i.e., after twelve transcripts) remained in this category during the entire analysis.

We showed above that high-frequency codes in the early stages of our analysis tended to retain their relative prevalence over time. But were there any high-frequency codes that emerged later in the analysis and that we would have missed had we only six or twelve interviews to analyze? The data in Table 5 address this question. After analyzing all sixty interviews, a total of thirty-six codes were applied with a high frequency to the transcripts. Of these, thirty-four (94%) had already been identified within the first six interviews, and thirty-five (97%) were identified after twelve. In terms of the range of commonly expressed themes, therefore, very little appears to have been missed in the early stages of analysis.

TABLE 5
Presence of High-Prevalence Codes in Early Stages of Analysis

Frequency after R10 (Sixty Interviews)	*Number of Codes*	*Percentage Present in R1 (First Six Interviews)*	*Percentage Present after R2 (First Twelve Interviews)*
High	36	94	97
Medium	39	56	83
Low	39	62	82

DISCUSSION

Based on our analysis, we posit that data saturation had for the most part occurred by the time we had analyzed twelve interviews. After twelve interviews, we had created 92% (100) of the total number of codes developed for all thirty of the Ghanaian transcripts (109) and 88% (114) of the total number of codes developed across two countries and sixty interviews. Moreover, four of the five new codes identified in the Nigerian data were not novel in substance but rather were variations on already existing themes. In short, after analysis of twelve interviews, new themes emerged infrequently and progressively so as analysis continued.

Code definitions were also fairly stable after the second round of analysis (twelve interviews), by which time 58% of all thirty-six definition revisions had occurred. Of the revisions, more than three-fourths clarified specifics and did not change the core meaning of the code. Variability of code frequency appears to be relatively stable by the twelfth interview as well, and, while it improved as more batches of interviews were added, the rate of increase was small and diminished over time.

It is hard to say how generalizable our findings might be. One source of comparison is consensus theory developed by Romney, Batchelder, and Weller (1986). Consensus theory is based on the principle that experts tend to agree more with each other (with respect to their particular domain of expertise) than do novices and uses a mathematical proof to make its case. Romney, Batchelder, and Weller found that small samples can be quite sufficient in providing complete and accurate information within a particular cultural context, as long as the participants possess a certain degree of expertise about the domain of inquiry ("cultural competence"). Romney, Batchelder, and Weller (1986:326) calculated that samples as small as four individuals can render extremely accurate information with a high confidence level (.999) if they possess a high degree of competence for the domain of inquiry

in question (1986:326). Johnson (1990) showed how consensus analysis can be used as a method for selecting participants for purposive samples.

While consensus theory uses structured questions and deals with knowledge, rather than experiences and perceptions per se, its assumptions and estimates are still relevant to open-ended questions that deal with perceptions and beliefs. The first assumption of the theory is that an external truth exists in the domain being studied, that there is a reality out there that individuals experience. Some might argue that in the case we presented, there is no external truth because we asked participants their opinions and perceptions, rather than, say, asking them to identify and name species of plants. This is partially true, but the individuals in our sample (and in most purposive samples/subsamples for that matter) share common experiences, and these experiences comprise truths. Many women in our study, for example, talked about fear of being exposed (i.e., their involvement in sex work) to the public, particularly via the media. Such fear and distrust is a reality in the daily lives of these women and is thus reflected in the data.

The second and third assumptions within the consensus model are that participants answer independently of one another and that the questions asked comprise a coherent domain of knowledge. The former assumption can be met by ensuring that participants are interviewed independently and in private. The latter assumption can be achieved by analyzing data collected from a given instrument compartmentally, by domain. Moreover, the data themselves can provide insights into the degree to which knowledge of one domain transfers to another. Themes that are identified across multiple domains and shared among numerous participants could be identified, post facto, as part of one larger "domain" of experience.

Our study included a relatively homogeneous population and had fairly narrow objectives. This brings up three related and important points: interview structure and content and participant homogeneity. With respect to the first point, we assume a certain degree of structure within interviews; that is, a similar set of questions would have to be asked of all participants. Otherwise, one could never achieve data saturation; it would be a moving target, as new responses are given to newly introduced questions. For this reason, our findings would not be applicable to unstructured and highly exploratory interview techniques.

With respect to instrument content, the more widely distributed a particular experience or domain of knowledge, the fewer the number of participants required to provide an understanding of the phenomenon of interest. You would not need many participants, for example, to find out the name of the local mayor or whether the local market is open on Sunday. Even a small convenience sample would likely render useful information in this case. Con-

versely, as Graves (2002:169) noted, "Lack of widespread agreement among respondents makes it impossible to specify the 'correct' cultural belief."

It really depends on how you want to use your data and what you want to achieve from your analysis. As Johnson (1998:153) reminds us, "It is critical to remember the connection between theory, design (including sampling), and data analysis from the beginning, because how the data were collected, both in terms of measurement and sampling, is directly related to how they can be analyzed." If the goal is to describe a shared perception, belief, or behavior among a relatively homogeneous group, then a sample of twelve will likely be sufficient, as it was in our study. But if one wishes to determine how two or more groups differ along a given dimension, then you would likely use a stratified sample of some sort (e.g., a quota sample) and might purposively select twelve participants per group of interest.

If your aim is to measure the degree of association between two or more variables using, say, a nonparametric statistic, you would need a larger sample. Graves (2002:72-75) presented an example of a two × two contingency table of height and weight of the San Francisco 49ers. Using a sample of 30, Graves calculated a chi-square value of 3.75 for the association between height and weight. This value is not quite statistically significant at the .05 level. However, when the sample size is doubled, but the relative proportions are kept constant, the chi-square value doubles to 7.5, which is highly significant. Graves (2002:73) therefore recommended collecting samples of between 60 and 120 for such correlative analyses (and, the larger the number, the more ways you can cross-cut your data).[3]

Our third point relates to sample homogeneity. We assume a certain degree of participant homogeneity because in purposive samples, participants are, by definition, chosen according to some common criteria. The more similar participants in a sample are in their experiences with respect to the research domain, the sooner we would expect to reach saturation. In our study, the participants were homogeneous in the sense that they were female sex workers from West African cities. These similarities appear to have been enough to render a fairly exhaustive data set within twelve interviews. Inclusion of the younger, campus-based women, however, did require creating a few new codes relatively late in the analysis process, which may signal that their lifestyles and experiences are somewhat distinct from their street- and brothel-based counterparts, but as mentioned earlier, these "new" codes were really just variations on existing themes. Structuring databases in a way that allows for a subgroup analysis and that can identify thematic variability within a sample is one way to assess the cohesiveness of a domain and its relationship to sample heterogeneity.

A final issue we wish to raise pertains to codebook structure and the age-old "lumper-splitter problem." Indeed, we have met qualitative researchers whose codebooks contain more than five hundred codes (each with values!). At the other extreme, a researcher may extract only four or five themes from a large qualitative data set. Clearly, the perception of saturation will differ between these two instances; as Morse (1995) pointed out, saturation can be an "elastic" concept. At the crux of the discussion is how and when we define themes and how we eventually plan to present our data. Ryan and Bernard (2003) noted that the problem of defining a theme has a long history, and many terms have been used to describe what we call themes. The authors go on, however, to define themes as "abstract (and often fuzzy) constructs that link . . . expressions found in text" and that "come in all shapes and sizes" (p. 87). Ultimately, themes should be able to be linked to data points; that is, one should be able to provide evidence of a given theme within the text being analyzed. In our view, codes are different from themes, in that the former are formal renderings of the latter. Codes are applied to the data (often electronically), whereas themes emerge from the data.

Ryan and Bernard (2004) asserted that when and how saturation is reached depends on several things: (1) the number and complexity of data, (2) investigator experience and fatigue, and (3) the number of analysts reviewing the data. In addition, some researchers warn that completing analysis too soon runs the risk of missing more in-depth and important content (Wilson and Hutchinson 1990:123). While true, we feel that conceptualizing saturation primarily as researcher dependent misses an important point: How many interviews or data points are enough to achieve one's research objectives given a set research team? Without a doubt, anyone can find, literally, an infinite number of ways to parse up and interpret even the smallest of qualitative data sets. At the other extreme, an analyst could gloss over a large data set and find nothing of interest. In this respect, saturation is reliant on researcher qualities and has no boundaries. The question we pose, however, frames the discussion differently and asks, "Given x analyst(s) qualities, y analytic strategy, and z objective(s), what is the fewest number of interviews needed to have a solid understanding of a given phenomenon?" Could we, for example, go back through our data and find new themes to add to the 114 existing ones? Sure we could, but if we used the same analysts and techniques and had the same analytic objectives, it is unlikely. The data are finite, and the stability of our codebook would bear this out if the original parameters remained constant in a reanalysis.

We have discussed codebook development while processing data, as would be expected in a grounded theory approach. But many codebook revisions are removed from the data collection process and consist of restructur-

ing (usually hierarchically) the relationships between codes after code definitions have been finalized and code application completed. This is true in our case; we first identified as many codes as we thought were relevant to our objectives, finalized the codebook, and then discussed overarching themes. The result of such a process is often a codebook that has several higher level metathemes that may or may not serve as parent codes to children codes. Such post hoc rearrangement does not affect saturation per se—since the range of thematic content in the codebook does not change—but it will likely influence how we think about and present our data. We should also point out that a lumper may identify only a few metathemes in the first place and never have enough codes to bother with ordering themes hierarchically or applying a data reduction technique.

Regardless of how one derives metathemes from a data set, if it is these overarching themes that are of primary interest to the researcher, saturation, for the purpose of data presentation and discussion, will likely occur earlier in the process than if more fine-grained themes are sought. Our postcoding data reduction and interpretation process rendered four metathemes. It is difficult to say, post facto, whether we would have had enough context to have derived these metathemes early on in the process, but in retrospect, looking at the metathemes and their constituent code frequencies, enough data existed after six interviews to support these four themes. The basic elements were there. The connections among the codes that eventually made up the overarching themes, however, may not have been apparent in the early stages of analysis, or we may have identified several other themes that dwindled in importance as transcripts were added and the analysis progressed. Nonetheless, the magic number of six interviews is consistent with Morse's (1994) (albeit unsubstantiated) recommendation for phenomenological studies. Similar evidence-based recommendations can be found for qualitative research in technology usability. Nielsen and Landauer (1993) created a mathematical model based on results of six different projects and demonstrated that six evaluators (participants) can uncover 80% of the major usability problems within a system, and that after about twelve evaluators, this diagnostic number tends to level off at around 90%.[4]

Our experiment documents thematic codebook development over the course of analyzing sixty interviews with female sex workers from two West African countries. Our analysis shows that the codebook we created was fairly complete and stable after only twelve interviews and remained so even after incorporating data from a second country. If we were more interested in high-level, overarching themes, our experiment suggests that a sample of six interviews may have been sufficient to enable development of meaningful themes and useful interpretations. We call on other researchers to conduct

similar experiments to see if, in fact, our results are generalizable to other domains of inquiry (particularly broader domains), types of groups, or other forms of data collection methods, such as focus groups, observation, or historical analysis.

At the same time, we want to caution against assuming that six to twelve interviews will always be enough to achieve a desired research objective or using the findings above to justify "quick and dirty" research. Purposive samples still need to be carefully selected, and twelve interviews will likely not be enough if a selected group is relatively heterogeneous, the data quality is poor, and the domain of inquiry is diffuse and/or vague. Likewise, you will need larger samples if your goal is to assess variation between distinct groups or correlation among variables. For most research enterprises, however, in which the aim is to understand common perceptions and experiences among a group of relatively homogeneous individuals, twelve interviews should suffice.

NOTES

1. Ethics review committees also usually require that sample sizes be written into protocols, and deviating from approved sampling procedures can involve time-consuming protocol amendments.

2. We chose sets of six because six is a divisor of thirty, and this number was the smallest recommended sample size we identified within the literature.

3. Note that although chi-square is highly useful for structured categorical responses, it is not suitable for data collected from an open-ended instrument such as ours. Contingency tables require that the frequencies in one cell are mutually exclusive and contrastive of other cells in the table (i.e., an individual weighs either 200 lb or more or 199 lb and less, or a medical intervention is either successful or not). In the case of open-ended questions, the presence of a trait within an individual (e.g., expression of a theme) cannot be meaningfully contrasted with the absence of this trait. That is, the fact that an individual does not mention something during the interview is not necessarily indicative of its absence or lack of importance.

4. Nielsen and Landauer (1993) also calculated that the highest return on investment was obtained with about five evaluators. It would be interesting to see if these monetary figures transfer to other domains of research.

REFERENCES

Bernard, H. R. 1995. *Research methods in anthropology*. Walnut Creek, CA: AltaMira.

———. 2000. *Social research methods*. Thousand Oaks, CA: Sage.

Bertaux, D. 1981. From the life-history approach to the transformation of sociological practice. In *Biography and society: The life history approach in the social sciences*, ed. by D. Bertaux, 29–45. London: Sage.

Bluff, R. 1997. Evaluating qualitative research. *British Journal of Midwifery* 5 (4): 232–35.

Brody, S. 1995. Patients misrepresenting their risk factors for AIDS. *International Journal of STD & AIDS* 6:392–98.

Byrne, M. 2001. Evaluating the findings of qualitative research. *Association of Operating Room Nurses Journal* 73 (3): 703–6.

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

Centers for Disease Control and Prevention. 2004. *AnSWR: Analysis Software for Word-based Records, version 6.4*. Atlanta, GA: Centers for Disease Control and Prevention.

Cheek, J. 2000. An untold story: Doing funded qualitative research. In *Handbook for qualitative research*, 2nd ed., ed. N. Denzin and Y. Lincoln, 401–20. Thousand Oaks, CA: Sage.

Creswell, J. 1998. *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Crosby, R., R. DiClemente, D. Holtgrave, and G. Wingood. 2002. Design, measurement, and analytical considerations for testing hypotheses relative to condom effectiveness against on-viral STIs. *Sexually Transmitted Infections* 78:228–31.

Flick, U. 1998. *An introduction to qualitative research: Theory, method and applications.* Thousand Oaks, CA: Sage.

Fossey, E., C. Harvey, F. McDermott, and L. Davidson. 2002. Understanding and evaluating qualitative research. *Australian and New Zealand Journal of Psychiatry* 36:717–32.

Geary, C., J. Tchupo, L. Johnson, C. Cheta, and T. Nyiama. 2003. Respondent perspectives on self-report measures of condom use. *AIDS Education and Prevention* 15 (6): 499–515.

Glaser, B. and A. Strauss. 1967. *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine Publishing Company.

Graves, T. 2002. *Behavioral anthropology: Toward an integrated science of human behavior*. Walnut Creek, CA: Roman & Littlefield.

Johnson, J. C. 1990. *Selecting ethnographic informants.* Thousand Oaks, CA: Sage.

———. 1998. Research design and research strategies. In *Handbook of methods in cultural anthropology*, ed. H. R. Bernard, 131–72. Walnut Creek, CA: AltaMira.

Kuzel, A. 1992. Sampling in qualitative inquiry. In *Doing qualitative research*, ed. B. Crabtree and W. Miller, 31–44. Newbury Park, CA: Sage.

LeCompte, M., and J. Schensul. 1999. *Designing and conducting ethnographic research.* Walnut Creek, CA: AltaMira.

MacQueen, K. M., E. McLellan, K. Kay, and B. Milstein. 1998. Codebook development for team-based qualitative analysis. *Cultural Anthropology Methods Journal* 10 (12): 31–36.

McLellan, E., K. M. MacQueen, and J. Niedig. 2003. Beyond the qualitative interview: Data preparation and transcription. *Field Methods* 15 (1): 63–84.

Miles, M., and A. Huberman. 1994. *Qualitative data analysis*. 2nd ed. Thousand Oaks, CA: Sage.

Morse, J. 1994. Designing funded qualitative research. In *Handbook for qualitative research*, ed. N. Denzin and Y. Lincoln, 220–35. Thousand Oaks, CA: Sage.

———. 1995. The significance of saturation. *Qualitative Health Research* 5:147–49.

Nielsen, J., and T. K. Landauer. 1993. A mathematical model of the finding of usability problems. *Proceedings of INTERCHI* 93:206–13.

Nunnally, J. C., and L. H Bernstein. 1994. *Psychometric theory*. 3rd ed. New York: McGraw-Hill.

Patton, M. 2002. *Qualitative research and evaluation methods.* 3rd ed. Thousand Oaks, CA: Sage.

Paulhus, D. 1991. Measurement and control of response bias. In *Measures of personality and social psychological attitudes*, Vol. 1, ed. J. Robinson, P. R. Shaver, and L. S. Wrightsman, 17–59. New York: Academic Press.

Romney, A., W. Batchelder, and S. Weller. 1986. Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist* 88:313–38.

Rubin, H., and I. Rubin. 1995. *Qualitative interviewing: The art of hearing data*. Thousand Oaks, CA: Sage.

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

———. 2004. Techniques to identify themes in qualitative data, http://www.analytictech.com/mb870/ryan-bernard_techniques_to_identify_themes_in.htm (accessed September 2004).

Sandelowski, M. 1995. Sample size in qualitative research. *Research in Nursing and Health* 18:179–83.

Schensul, S., J. Schensul, and M. LeCompte. 1999. *Essential ethnographic methods*. Walnut Creek, CA: AltaMira.

Schwarz, N. 1999. Self-reports: How the questions shape the answers. *American Psychologist* 54:93–105.

Trotter, R., II. 1991. Ethnographic research methods for applied medical anthropology. In *Training manual in applied medical anthropology*, ed. C. Hill, 180–212. Washington, DC: American Anthropological Association.

Trotter, R., II, and J. Schensul. 1998. Methods in applied anthropology. In *Handbook of methods in cultural anthropology*, ed. H. R. Bernard, 691–736. Walnut Creek, CA: AltaMira.

Weinhardt, M., A. Forsyth, M. Carey, B. Jaworski, and L. Durant. 1998. Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. *Archives of Sexual Behavior* 27:155–80.

Weir, S., R. Roddy, L. Zekeng, and K. Ryan. 1999. Association between condom use and HIV infection: A randomised study of self reported condom use measures. *Journal of Epidemiological Community Health* 53:417–22.

Wilson, H. S., and S. Hutchinson. 1990. Methodologic mistakes in grounded theory. *Nursing Research* 45 (2): 122–24.

Zenilman, J., C. Weisman, A. Rompalo, N. Ellish, D. Upchurch, E. Hook, and D. Celentano. 1995. Condom use to prevent STDs: The validity of self-reported condom use. *Sexually Transmitted Diseases* 22:15–21.

*Greg Guest is a sociobehavioral scientist at Family Health International, where he conducts research on the sociobehavioral aspects of reproductive health. His most recent work deals with HIV/AIDS prevention and behavioral components of clinical trials in Africa. Dr. Guest also has an ongoing interest in the ecological dimensions of international health and the integration of qualitative and quantitative methodology. His most recent publications include "Fear, Hope, and Social Desirability Bias Among Women at High Risk for HIV in West Africa" (*Journal of Family Planning and Reproductive Health Care*, forthcoming), "HIV Vaccine Efficacy Trial Participation: Men-Who-Have-Sex-With-Men's Experience of Risk Reduction Counseling and Perceptions of Behavior Change" (2005,* AIDS Care*), and the edited volume* Globalization, Health and the Environment: An Integrated Perspective *(2005, AltaMira). He is currently co-editing (with Kathleen MacQueen) the* Handbook for Team-Based Qualitative Research.

Arwen Bunce is a senior research analyst and qualitative specialist at Family Health International in North Carolina. Her research interests include the intersection of reproductive health and human rights, and the impact of sociocultural factors on women's health and well-being. Previous research experience include research surrounding access to medical care for immigrants and the construct of self-rated health. Her publications include "The Assessment of Immigration Status in Health Research" (with S. Loue, Vital Health Statistics*, 1999) and "The Effect of Immigration and Welfare Reform Legislation on Immigrants' Access to Health Care, Cuyahoga and Lorain Counties" (with S. Loue and M. Faust,* Journal of Immigrant Health*, 2000).*

Laura Johnson is a research associate at Family Health International. She performs qualitative and quantitative data analysis on a variety of research topics including: youth and family planning, reliability of self-reported data, and costs and delivery of family planning services. Her recent publications include "Respondent Perspectives on Self-Report Measures of Condom Use" (with C. Waszak Geary, J.P. Tchupo, C. Cheta, and T. Nyama, AIDS Education and Prevention, *2003), "Excess Capacity and the Cost of Adding Services at Family Planning Clinics in Zimbabwe" (with B. Janowitz, A. Thompson A, C. West, C. Marangwanda, and N.B. Maggwa,* International Family Planning Perspectives, *2002) and "Quality of Care in Family Planning Clinics in Jamaica: Do Clients and Providers Agree?" (with K. Hardee, O.P. McDonald OP, and C. McFarlane,* West Indian Medical Journal, *2001).*