RESEARCH NOTE

An Examination of the Sufficiency of Small Qualitative Samples

Diane S. Young and Erin A. Casey

Qualitative researchers often must make decisions about anticipated sample sizes in advance of data collection. Estimates are typically required for human subjects review committees, grant applications, and resource planning purposes. Once a study is underway or completed, researchers must evaluate whether the sample has been robust enough to address the research aims. The challenge is to find a sample that will produce thorough and meaningful findings while minimizing unnecessary burden on participants and expenditure of scarce resources such as time and research dollars. Currently, little guidance is available regarding what minimum sample size is needed to adequately identify the themes and codes in an area of inquiry. In addition, the issue of sample sizes needed to reach theme and code saturation across different qualitative methodologies or data analysis approaches is understudied.

Although researchers often cite having achieved saturation as a reason to conclude sampling, details regarding how saturation was determined are not provided for the most part (Bowen, 2008; Francis et al., 2010). Gentles, Charles, Ploeg, and McKibbon (2015) conducted an overview of the literature from influential authors within the traditions of grounded theory, phenomenology, and case study, and they noted the lack of clarity relative to sample size and saturation. Guetterman (2015) looked at the most-cited empirical articles in the fields of education and health sciences from 2008 through 2012 within five qualitative research approaches to assess specific samples sizes and the rationale for sample sizes. Sample size across the 51 studies varied widely, and most articles did not include a discussion of saturation or the adequacy of the sample.

In an effort to provide empirically based guidance about appropriate minimum sample sizes for qualitative studies, researchers have recently begun to conduct methodological studies that examine the point at which data saturation occurs. Guest, Bunce, and Johnson (2006) operationalized data saturation "as the point in data collection and analysis when new information produces little or no change to the codebook" (p. 65). They reviewed transcripts from a previous study in sets of six, according to the order in which individual interviews had been conducted at two research sites. They noted theme and code development, asking whether six interviews yielded as much data as 12, 18, 24, and so on interviews. They found that 73% of codes were identified in the first six interviews and 92% within the first 12 interviews. Examining this same question with focus group data, as opposed to individual interviews as in the earlier study, Guest, Namey, and McKenna (2016) found that 60% of their 94 codes were found in the first focus group, 84% in the first three groups, and 90% by six. When the focus groups were randomly ordered to assess for temporal bias, the results remained consistent.

Other researchers have examined the question of minimum sample size using different definitions of data saturation, sometimes referred to as code saturation. Extending Guest et al.'s (2006) findings to cross-cultural research, Hagaman and Wutich (2016) considered three repetitions of a theme by different interviewees as identification of that theme and found that 16 interviews were enough to identify themes from homogeneous groups, with 20 to 40 needed to identify metathemes that cut across cultures and study sites. Francis and colleagues (2010) considered data saturation to be achieved when no new ideas emerged with three additional interviews. Using this strict stopping criterion of three, they examined interview data from two different studies and found that saturation was achieved in one study at 17 interviews, with no new data emerging after the 14th interview, and was not yet

doi: 10.1093/swr/svy026

determined in the 14 interviews available to them from the second study. Even so, the majority of themes (92% and 86% in study one and two, respectively) emerged in the first six interviews. Finally, Hennink, Kaiser, and Marconi (2016) proposed that code saturation is the point at which you have "heard it all," but that meaning saturation is the point when you "understand it all" (p. 15). In their study on patient retention with 25 individuals, they found that 84% of codes were identified by the sixth interview and 91% by the ninth interview. It took 16 to 24 interviews, however, to understand all the dimensions of the nine central codes, achieving meaning saturation.

These findings suggest that under some study conditions, rich qualitative findings can be discovered with relatively small sample sizes. Further determining the parameters under which this applies would be helpful to researchers and research participants alike. Most efforts thus far have been done with studies relying on individual interviews, and many are within the medical field. In addition, examinations of minimal required sample sizes that examine available interviews once, in the order they were collected, raise concerns about possible temporal bias. We sought to examine the minimum sample sizes needed to adequately include the themes and codes in areas of inquiry within the field of social work. Considering three distinct qualitative research studies inclusive of both individual interviewing and focus group data collection approaches, we addressed four research questions: (1) What minimum sample size is needed to adequately identify codes (smaller units of meaning) within the data? (2) What minimum sample size is needed to ensure that all larger themes are partially represented by at least one of the codes that comprise that theme? (3) What minimum sample size is needed to fully realize the complete dimensionality of all themes by including all assigned codes? (4) Are minimum sample sizes needed consistent across different substantive areas of exploration and different modes of data collection, specifically individual interviews and focus groups? To address temporal bias, we addressed these questions by examining multiple random draws of various sample sizes within each included qualitative study.

METHOD

For the purpose of addressing the stated questions related to sample size and data redundancy, this article presents analyses done on data we previously collected for three distinct qualitative studies. Each original study is described briefly, outlining each one's research aims, sample size and participant criteria, mode of data collection, analytic process, and number of resulting themes and codes. These brief synopses are presented to indicate the diversity of substantive areas and approaches used. More detail about each, including original research findings, are referenced. For the present methodological study, data from the original studies were not reanalyzed. Rather, the presence or absence of the themes and codes originally identified and described in the cited, published studies were examined in random subsamples.

The Men Against Violence Study

The Men Against Violence (MAV) study (Casey, 2010) consisted of individual interviews with 27 U.S. men between the ages of 20 and 72 who identified as allies in the prevention of gender-based violence. The primary aim of the study was to assess the strategies used and challenges faced by the participants as they work to engage other men and boys in violence prevention. Respondents represented all regions of the United States and were recruited via topic-relevant Listservs and referrals from violence prevention organizations. Data were gathered in person or over the phone via a uniform, semistructured interview guide that assessed the nature of men's antiviolence involvement, their use and perceptions of effective and ineffective strategies for engaging other men, and the barriers they encountered in efforts to reach men.

Once all interviews were conducted, transcripts resulting from the interviews were analyzed using techniques drawn from grounded theory and described by Charmaz (2006). Analysis included inductive, line-by-line coding of transcript content, in conjunction with extensive author memo making to uncover concepts within the data. Axial coding then used a constant comparative method both within and between cases to identify larger themes from a finalized list of more specific codes. This process identified four themes comprising 20 codes, or more specific units of meaning that collectively defined the full dimensionality of each theme.

The Social Workers in Criminal Justice Study

The Social Workers in Criminal Justice (SWCJ) study (Young, 2014) consisted of individual interviews with 15 experienced social workers working within diverse criminal justice settings in the northwestern United States. Participants shared their perspectives about the definitions of success and attributes needed for effective social work practice in their roles within adult prison, juvenile rehabilitation, treatment court, and offices of prosecution and public defense. Snowball sampling was used to locate individuals with an undergraduate or graduate degree in social work and currently practicing social work in a criminal justice setting. Interviews were conducted in person or over the phone with the use of a semistructured, uniform interview guide.

Description rather than theory building shaped the analysis approach. Coding categories were gleaned from the text in relation to the general open-ended research questions: "How do you define success in your work?" and "What personal attributes are needed to be successful in your line of work?" The transcripts in their entirety were reviewed after all interviews were conducted. Once coding categories were identified and all transcripts were coded, the list of initial codes was reviewed and placed into conceptual groupings of major themes and subthemes. Then another thorough review of the transcripts was done, applying the revised set of coding categories to the transcripts and double-checking that the final set of themes and subthemes captured the ideas of the participants. This process identified eight themes comprising 30 specific units of meaning (codes) that collectively defined the full dimensionality of the themes.

The Adolescent Bystander Behavior Study

The Adolescent Bystander Behavior (ABB) study (Casey, Lindhorst, & Storer, 2017) aimed to identify influences on adolescent bystander decision making in the context of dating violence and bullying. More specifically, the project examined the relevance of two specific behavioral theories (the situational model of bystander behavior and the theory of planned behavior) to explaining bystander behavior. Data were gathered through 12 focus groups with a total of 113 youths ages 14 to 18; eight of these were face-to-face focus groups in local high schools and youth-serving agencies, and four groups were conducted in a real-time online format via text-based chat. Focus groups were facilitated by two researchers and data were gathered using a semistructured, uniform interview guide. Youths were asked to identify common dating violence and bullying scenarios, and then to talk in depth about the range of factors that would influence their decision making regarding how they might respond to these scenarios as bystanders.

Data analysis proceeded in two phases once interviews were finished. First, deductive coding (Miles, Huberman, & Saldaña, 2014) was used to identify content in the transcripts relevant to the five constructs that collectively comprise the two guiding theoretical frameworks. Once all the transcripts were analyzed for content relevant to larger theory constructs, inductive thematic content analysis was used to identify codes reflecting the beliefs and ideas that collectively defined each larger theory construct. In addition, content regarding influences on bystander decision making that was not contained within the guiding theories was also inductively coded. These processes resulted in seven larger themes (the five theory constructs and two additional themes), which were defined by a total of 37 codes.

The Present Methodological Study

In our study, we retrospectively used the data and findings from the three previously described projects because they have important similarities and differences critical to addressing our research aims. Across the projects, interview or focus group data were transcribed and the transcripts thoroughly analyzed, resulting in a specific number of relevant themes and codes. However, in each project researchers addressed different topics and collectively gathered data through two methods: individual interviews and focus groups. Table 1 provides a listing of the number of cases, themes, and codes present in the

Table 1: Sample Size and Number of Themes and Codes in the Reviewed Studies

Study Name (Data Collection Method)	Cases (*n*)	Themes (*n*)	Codes (*n*)
MAV (individual interviews)	27	4	20
SWCJ (individual interviews)	15	8	30
ABB (focus groups)	12	7	37

Notes: MAV = Men Against Violence; SWCJ = Social Workers in Criminal Justice; ABB = Adolescent Bystander Behavior.

original studies. Each individual interview or focus group transcript represents a case.

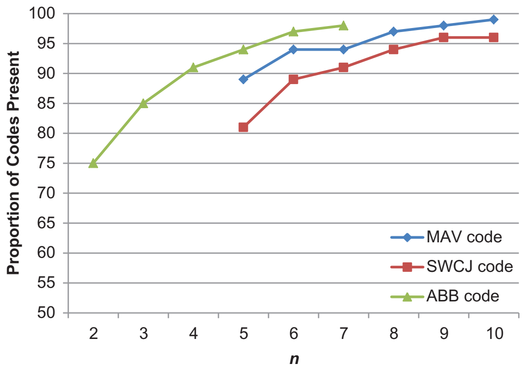
A data set for each original study was created that identified for each transcript the presence or absence of the previously determined themes and codes. Then, using a random number generator, 10 random samples of each size from $n = 5$ through $n = 10$ for individual interviews and $n = 2$ through $n = 7$ for focus groups were drawn from each project. Because one focus group potentially yields more information than one individual interview and the research aims sought to determine minimum sample sizes, the number of focus groups comprising the subsamples was adjusted downward. To address the research aims, each randomly drawn subsample was then examined to see what proportion of the codes and larger themes from each original study's full sample were present within each subsample. Finally, results from all 10 subsamples for a given sample size were averaged together to determine the mean presence (expressed as a percentage) of codes and themes.

RESULTS

The first research aim was to examine at what sample size all final codes within the data were, on average, represented in the randomly selected transcripts. For interview-based projects, near-complete representation of codes was achieved at $n = 8$ in the MAV project (with an average of 97% of codes represented across random draws), and $n = 9$ in the SWCJ project (96% of codes represented). Adding one additional transcript to these sample sizes increased representation only to 98% in the MAV project, and did not add new coverage in the SWCJ project. The ABB focus group project achieved near-perfect code coverage at a sample size of six focus groups, with an average of 97% of codes represented across the random draws. Increasing the sample size to seven only increased coverage to an average of 98% of all possible codes. Near total inclusion of codes thus varied between $n = 6$ and $n = 9$ across the three qualitative projects. No project evidenced 100% average coverage across all draws at any sample size. Some individual draws reached 100% coverage starting at $n = 5$ for the MAV and ABB focus group projects, and $n = 8$ for the SWCJ project. Average code coverage findings are graphed in Figure 1.

The second aim sought to identify the sample size at which all larger themes were at least partially represented by one or more codes within each theme. Findings show that at least some aspect of all larger themes are present at sample sizes ranging from 4 to 6. More specifically, the MAV and SWCJ projects reached 100% average partial representation of themes at $n = 5$ and $n = 6$, respectively. The ABB focus group project reached consistent partial theme representation at $n = 4$.

Figure 1: Average Proportion of Codes Present in Each Set of Random Samples of *n* Transcripts

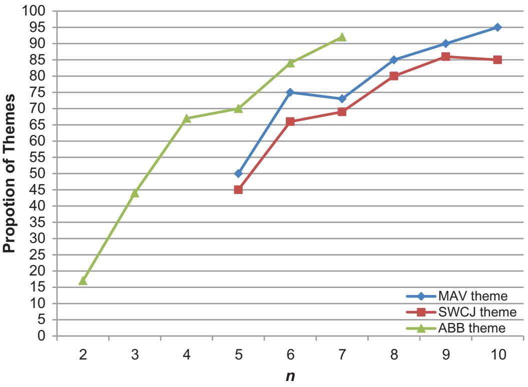


Notes: MAV = Men Against Violence; SWCJ = Social Workers in Criminal Justice; ABB = Adolescent Bystander Behavior.

Our third aim was to assess the sample sizes at which themes are fully realized within the data, that is, the point at which themes are defined by the full complement of codes that comprise them. These findings are presented in Figure 2. None of the three projects reached 100% theme completion at any of the examined sample sizes, although the percentage of fully defined themes was relatively high even with small samples. Specifically, the MAV project demonstrated 90% and 95% average theme realization at sample sizes of $n = 9$ and $n = 10$, respectively. The SWCJ project showed slightly lower theme completion with 86% average coverage at $n = 9$ and 85% average coverage at $n = 10$. For the ABB focus group data, 84% of themes were fully realized, on average, at $n = 6$, and 92% were completed at a sample size of seven. On some individual draws, however, 100% theme realization was found at $n = 5$ on the MAV project, $n = 6$ on the SWCJ project, and $n = 5$ on the ABB focus group project.

Relative to our research aim 4, code and theme representation occurred at similar sample sizes within the three projects examined here across all metrics. As previously summarized, significant coverage of codes ranged from a minimum sample size of six to nine,

Figure 2: Average Proportion of Fully Realized Themes in Each Set of Random Samples of *n* Transcripts



Notes: MAV = Men Against Violence; SWCJ = Social Workers in Criminal Justice; ABB = Adolescent Bystander Behavior.

partial theme representation required minimum sample sizes of four to six, and substantial theme completion necessitated sample sizes of seven to 10 cases across the projects. The ABB focus group project was consistently at the lower end of these ranges, and the more code-heavy of the individual interview projects (SWCJ) typically occupied the higher end.

DISCUSSION

In three substantive areas, using two methodologies frequently used in qualitative research, findings from small subsamples adequately identified themes and codes in each area of inquiry. These findings agree with previous research (Guest et al., 2006, 2016; Hennink et al., 2016) and provide an important replication and extension of others' work. The question about what sample size is sufficient is a critical methodological one, affecting almost all qualitative researchers. These findings give strong evidence and reassurance that researchers, under certain conditions, can achieve robust results with small sample sizes. Doing so will minimize participant burden and maximize limited resources.

Clarifying the conditions under which small sample sizes yield meaningful findings will further benefit fields that rely heavily on qualitative research approaches. This is an important focus for future research. Aspects of the studies we drew on incorporated factors that are thought to contribute to the ability to achieve thorough findings with small sample sizes: participants met predetermined criteria and described similar experiences, and interviews were relatively structured (Guest et al., 2006; Malterud, Siersma, & Guassora, 2016). Extending the methods previously included in similar examinations of sample size and data redundancy, the studies we drew on incorporated in-person and telephone individual interviews and in-person and real-time online focus groups. That findings were consistent regardless of data collection method strengthens the conclusion that small qualitative samples are adequate for producing robust findings. In guarding against temporal bias by randomly drawing subsamples, we also found that the order in which the transcripts were examined was important. As few as five transcripts included all codes (100%) in some of the individual random sample draws for two out of the three research projects. Using randomization of multiple sample draws helped mitigate against conclusions based on early outliers. This may be a useful approach to continue in future studies.

Our findings contribute to the growing body of evidence that robust identification of themes and codes may be achieved relatively quickly in interview and focus group data. Additional cases rounded out or added slight nuance to identified themes, but the vast majority of codes and themes were present in small samples. These findings echo conclusions reached by Hennink et al. (2016), who found near code saturation ("hearing it all") at six to nine interviews, and additional nuance ("understanding it all") as additional transcripts were included. The accumulating evidence across studies therefore suggests that rigorously collected qualitative data from small samples can substantially represent the full dimensionality of people's experiences, with larger sample sizes adding important but perhaps increasingly minute pieces of meaning. Small sample size should not be seen as a limitation, in and of itself, when evaluating the rigor and findings of qualitative research. **SWR**

REFERENCES

Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative Research, 8*(1), 137–152.

Casey, E. A. (2010). Strategies for engaging men as anti-violence allies: Implications for ally movements. *Advances in Social Work, 11*, 267–282.

Casey, E. A., Lindhorst, T. P., & Storer, H. L. (2017). The situational-cognitive model of adolescent bystander behavior: Modelling bystander decision making in the context of bullying and teen dating violence. *Psychology of Violence, 7*, 33–44.

Charmaz, K. (2006) *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA: Sage Publications.

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology and Health, 25*, 1229–1245.

Gentles, S. J., Charles, C., Ploeg, J., & McKibbon, K. A. (2015). Sampling in qualitative research: Insights from an overview of the methods literature. *Qualitative Report, 20*, 1772–1789.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59–82.

Guest, G., Namey, E., & McKenna, K. (2016). How many focus groups are enough? Building an evidence base for nonprobability sample sizes. *Field Methods, 29*(1), 3–22.

Guetterman, T. C. (2015). Descriptions of sampling practices within five approaches to qualitative research in education and the health sciences. *Forum: Qualitative Social Research, 16*(2), Article 25.

Hagaman, A. K., & Wutich, A. (2016). How many interviews are enough to identify metathemes in multisited and cross-cultural research? Another perspective on Guest, Bunce, and Johnson's (2006) landmark study. *Field Methods, 29*(1), 23–41.

Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2016). Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research, 27*, 1–18.

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research, 26*, 1753–1760.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Young, D. S. (2014). Social workers' perspectives on effective practice in criminal justice settings. *Journal of Forensic Social Work, 4*(2), 104–122.

Diane S. Young, PhD, MSW, *is director and associate professor and* ***Erin A. Casey, PhD, MSW,*** *is professor, Social Work and Criminal Justice Program, University of Washington–Tacoma. Address correspondence to Diane S. Young, Social Work and Criminal Justice Program, University of Washington–Tacoma, 1900 Commerce Street, Box 358425, Tacoma, WA 98402; e-mail: youngd4@uw.edu.*

Original manuscript received March 1, 2017
Final revision received June 22, 2017
Editorial decision July 6, 2017
Accepted July 7, 2017