Determining Appropriate Sample Size for Qualitative Interviews: Code Saturation

Claudia M Squire, Kristen C Giombi, Douglas J Rupert, Jacqueline Amoozegar, Peyton Williams

Submitted to: Journal of Medical Internet Research
on: September 22, 2023

Table of Contents

Original Manuscript **5**
Supplementary Files **20**
0 ..... **20**
0 ..... **20**
0 ..... **20**
Multimedia Appendixes **21**
Multimedia Appendix 0 ..... **21**

Determining Appropriate Sample Size for Qualitative Interviews: Code Saturation

Claudia M Squire[1] MS; Kristen C Giombi[1] PhD; Douglas J Rupert[1] MPH; Jacqueline Amoozegar[1] MSPH; Peyton Williams[1] MPH

[1]RTI International Research Triangle Park US

Corresponding Author:
Claudia M Squire MS
RTI International
3040 E. Cornwallis Road
Research Triangle Park
US

Abstract

Background: In-depth interviews are a common method of qualitative data collection that allows for gathering rich data on individuals' perceptions and behaviors that would be challenging to collect with quantitative methods. Researchers typically need to make decisions on sample size a priori. Although studies have assessed when saturation has been achieved, there is no agreement on the minimum number of interviews needed to achieve saturation. Additionally, to date most research on saturation has been based on in-person data collection.

Objective: This study aimed to identify the number of virtual interviews needed to achieve true code saturation or near code saturation.

Methods: The analyses for this study were based on data from 5 Food and Drug Administration-funded studies conducted virtually with patients with underlying medical conditions or with healthcare providers who provide primary or specialty care to patients. We extracted code- and interview-specific data and examined the data summaries to determine when true saturation or near saturation was reached.

Results: The sample size used in the 5 studies ranged from 30 to 70 interviews. True saturation was reached after 91% to 100% of planned interviews, whereas near saturation was reached after 33% to 60% of planned interviews (15-23 interviews). Studies that relied heavily on deductive coding and studies that had a more structured interview guide reached both true saturation and near saturation sooner.

Conclusions: Our study provides support that near saturation may be a sufficient measure to target and that conducting additional interviews after that point may result in diminishing returns. Factors to consider in determining how many interviews to conduct include the structure and type of questions included in the interview guide, the coding structure, and the population under study. Studies with less structured interview guides, studies that rely heavily on inductive coding and analytic techniques, and studies that include populations that may be less knowledgeable about the topics discussed may require a larger sample size to reach an acceptable level of saturation.

(JMIR Preprints 22/09/2023:52998)
DOI: https://doi.org/10.2196/preprints.52998

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Determining Appropriate Sample Size for Qualitative Interviews: Code Saturation

Abstract

Background: In-depth interviews are a common method of qualitative data collection that allows for gathering rich data on individuals' perceptions and behaviors that would be challenging to collect with quantitative methods. Researchers typically need to decide on sample size a priori. Although studies have assessed when saturation has been achieved, there is no agreement on the minimum number of interviews needed to achieve saturation. To date, most research on saturation has been based on in-person data collection. During the COVID-19 pandemic, virtual data collection became increasingly common as traditional in-person data collection was possible. Researchers continue to use virtual data collection methods post-COVID, making it important to assess whether findings around saturation differ for in-person versus virtual interviews.

Objective: This study aimed to identify the number of virtual interviews needed to achieve true code saturation or near code saturation.

Methods: The analyses for this study were based on data from 5 Food and Drug Administration-funded studies conducted virtually with patients with underlying medical conditions or with healthcare providers who provide primary or specialty care to patients. We extracted code- and interview-specific data and examined the data summaries to determine when true saturation or near saturation was reached.

Results: The sample size used in the 5 studies ranged from 30 to 70 interviews. True saturation was reached after 91% to 100% of planned interviews, whereas near saturation was reached after 33% to 60% of planned interviews (15-23 interviews). Studies that relied heavily on deductive coding and studies that had a more structured interview guide reached both true saturation and near saturation sooner. We also examined the types of codes applied after near saturation had been reached. In 4 of the 5 studies, most of these codes represented previously established core concepts or themes. Codes representing newly identified concepts, other/miscellaneous responses (e.g., "in general"), uncertainty or confusion (e.g., "don't know"), or categorization for analysis (e.g., correct as compared with incorrect) were less commonly applied after near saturation had been reached.

Conclusions: Our study provides support that near saturation may be a sufficient measure to target and that conducting additional interviews after that point may result in diminishing returns. Factors to consider in determining how many interviews to conduct include the structure and type of questions included in the interview guide, the coding structure, and the population under study. Studies with less structured interview guides, studies that rely heavily on inductive coding and analytic techniques, and studies that include populations that may be less knowledgeable about the topics discussed may require a larger sample size to reach an acceptable level of saturation. Our findings also build on previous studies looking at saturation for in-person data collection conducted at a small number of sites.

Keywords: Saturation, sample size, virtual data collection, semistructured interviews

Introduction

In-depth interviews are commonly used to collect qualitative data for a wide variety of research

purposes across many subject matter areas. These types of interviews are an ideal approach for examining individuals' perceptions and behaviors at a level of depth, complexity, and richness that would be challenging to achieve with quantitative data collection methods. Typically, trained interviewers conduct interviews using a guide designed to address the study's key research aims by asking a series of questions and probes ordered by topic. These interview guides can range from highly structured to completely unstructured (e.g., loosely organized conversations). Following the completion of data collection, interview notes and transcripts generated from audio-recordings of the interviews are analyzed to assess for patterns in responses among the interviewees or subsets of the participants. [1, 2]

During the COVID-19 pandemic, virtual data collection became increasingly common as traditional in-person data collection was not possible, and researchers continue to use virtual data collection methods post-COVID, citing advantages such as accessing marginalized populations, achieving greater geographic diversity, being able to offer a more flexible schedule, and saving on travel expenses.[3] Potential concerns about virtual data collection, such as the inability to build rapport and data richness, have been largely unfounded.[3, 4]

While we do not expect virtual data collection to supplant in-person research, it continues to show signs of growth. To date, much of the research on qualitative methods has focused on in-person data collection. Consequently, it will be important to conduct research to determine if previous widely accepted findings hold true for virtual data collection.

Researchers typically make a priori decisions about the number of interviews to conduct with the aim of balancing the need for sufficient data with resource limitations and respondent burden. The concept of saturation is frequently used to justify the study's rigor with respect to the selected sample size. To provide empirically based recommendations on adequate minimum sample sizes, researchers have conducted studies to assess when saturation occurs. However, multiple types of saturation exist—such as, theoretical, thematic, code, and meaning—and within each type of saturation the definitions and measurement approaches used by investigators vary substantially, as does the level of detail researchers report in publications about their methods for achieving and assessing saturation. [5]

The present study aimed to examine the number of interviews needed to obtain code saturation for 5 recently conducted studies funded by the Food and Drug Administration[6] (FDA) involving virtual interviews. Specifically, how many virtual interviews are needed to obtain *true* code saturation (i.e., the use of 100% of all codes applied in the study) and how many virtual interviews are needed to achieve *near* code saturation (i.e., the use of 90% of all codes applied in the study)?

Literature Review

Multiple authors have defined saturation as the point during data collection and analysis at which no new additional data are found that reveal a new conceptual category[7-13] or theme related to the research question—an indicator that further data collection is redundant. [11] Additionally, Coenen et al. specified that no new second-level themes are revealed in 2 consecutive focus groups or interviews. [14]

Other authors have distinguished between various types of saturation. One of the most common types of saturation mentioned in the literature is theoretical saturation, which emerges from grounded theory and occurs when the concepts of a theory are fully reflected in the data and no new insights, themes, or issues are identified from the data. [5, 11, 12, 15-18] Henninck et al. [17] expanded this definition, adding that all relevant conceptual categories should have been identified, thus emphasizing the importance of sample adequacy over sample size. Guest et al.[19] operationalized the concept of theoretical saturation as the point in data collection and analysis when new information produces little or no change to the codebook,

and van Rijnsoever operationalized it as being when all the codes in the population have been observed once in the sample. [20]

Some authors have defined theoretical saturation, thematic saturation, and data saturation as the same concept, [16, 18] whereas others have defined these terms differently. [12, 21] For example, some authors have defined thematic saturation as the point where no new codes or themes are emerging from the data. [12, 22] For thematic saturation to be achieved, data should be collected until nothing new is generated. [21, 23] Data saturation has been defined as the level to which new data are repetitive of the data that have been collected. [12, 24, 25]

Furthermore, Henninck et al distinguished between code saturation and meaning saturation. [17] Code saturation is based on primary/parent codes and relates to the quantity of the data ("hearing it all"). Meaning saturation is based on sub/child codes and relates to the quality or richness of the data ("understanding it all"). Constantinou et al. [7] made the point that it is the categorization of the raw data, rather than the data, that is saturated.

The literature reflects multiple methods that have been used to determine saturation. [7-10, 13-18, 22, 26] Sim et al. [27] discussed the 4 general approaches that have been used to determine sample size for qualitative research: (1) rules of thumb, based on a combination of methodological considerations and past experience; (2) conceptual models, based on specific characteristics of the proposed study; (3) numerical guidelines derived from empirical investigation; and (4) statistical approaches, based on the probability of obtaining a sufficient sample size.

For example, Galvin[9] used a statistical approach based on binomial logic to establish the relationship between identifying a theme in a particular sample and within the larger population; for example, n% chance of detecting a theme if that theme exists within n% of the population. Using the probability equation, the researcher can determine the number of interviews needed for a stated level of confidence that all relevant themes held by a certain proportion of the population will occur within the interview sample. This method assumes the researcher knows in advance the emergent themes from the study and at what rate they may occur.

Constantinou et al.[7] used the Comparative Method for Themes Saturation, which relies on both a deductive and an inductive approach to generate codes (keywords extracted from the participants' words) and themes (codes that fall into similar categories). Themes are compared across interviews and theme saturation is reached when the next interview does not produce any new themes. The sequence of interviews is reordered multiple times to check for order-induced error.

When exploring the various methods for determining saturation, researchers reached different conclusions on when saturation was achieved (Table 1).

Table 1. Achieving saturation in interviews: Saturation type, methods for achieving saturation and findings by other authors

Citation	Saturation type	Method for achieving saturation	Finding
Coenen et al. [14]	Data saturation	Compared 2 approaches to conducting focus groups and individual interviews: an open approach and an approach based on the International Classification of Functioning, Disability and Health (ICF)	• Open approach (open-ended questions were used): saturation reached after 9 interviews • ICF-based approach (first level of the ICF classification was added to the open-ended questions): saturation

			reached after 12 interviews
Constantinou C [7]	Thematic saturation	Used Comparative Method for Themes Saturation (CoMeTS)	• Theme saturation reached at interview 5 • After reordering interviews 3 different ways, theme saturation reached at the 7th, 8th, and 8th interviews
Francis JJ et al. [8]	Data saturation	Two steps: 1. Specified a priori a minimum sample size for initial analysis, which depends on the complexity of the research questions/ interview guide, the diversity of the sample, and the nature of the analysis 2. Specified a priori how many more interviews will be conducted without new ideas emerging (stopping criterion)	• After 10 interviews (the initial analysis sample), 57 shared beliefs identified; no new shared beliefs in interviews 11 or 12; 2 new shared beliefs at interview 13; applying the stopping criterion indicates that study-wise saturation was not achieved in Study 1 • Saturation was achieved after 17 interviews in Study 2
Fugard AJ, Potts HW [26]	Thematic saturation	Conducted a statistical calculation of saturation based on expected theme prevalence within the population, number of desired instances of the theme, and desired power of the study	• To have 80% power to detect 2 instances of a theme with 10% prevalence, 29 interviews are required
Galvin R [9]	Thematic saturation	Employed a statistical approach, based on binomial logic, to ascertain the relationship between theme identification in a particular sample and the larger population	• If the researcher needs to be at least 95% confident that all the issues have emerged that are represented in 10% or more of the population, then 29 interviews are required
Guest G, Bunce A [15]	Thematic saturation	Operationalized saturation as a proportion: the number of identified themes at a given point in analysis divided by the total number of themes identified in the entire sample; level of saturation reported as the point at which, post facto, 80% or 90% of themes in a dataset are identified	• Saturation reached after 12 interviews; basic elements for meta-themes were present as early as 6 interviews

Guest G, Namey E, Chen M. [10]	Thematic saturation	3 elements: base size, run length, and new information threshold • Base size: the minimum number of interviews that should be reviewed/ analyzed to calculate the amount of information already gained • Run length: the number of interviews within which we look for and calculate new information; the number of new themes found in the run defines the numerator in the saturation ratio • New information threshold: what level of paucity of new information should we accept as indicative of saturation?	• Reached <5% new information threshold at 6 interviews across all base sizes with a run length of 2 interviews • Reached <5% new information threshold at 7 interviews across all base sizes with a run length of 3 interviews • Reached 0% new information threshold at 11 interviews across all base sizes with a run length of 2 interviews • Reached 0% new information threshold at 14 interviews across all base sizes with a run length of 3 interviews
Hagaman A K, Wutich A [16]	Thematic saturation	Used Ryan and Bernard's (2003) repetition approach to identify themes	• Top 3 themes identified at least once in as few as 5 interviews; took more than 24 interviews to elicit the top 3 themes from 3 different respondents • 9 meta-themes identified at least once in as few as 8 interviews; took more than 39 interviews to elicit the 9 meta-themes from 3 different respondents • 16 or fewer interviews needed to identify common themes from a fairly homogenous group
Hennink MM, Kaiser BN, Marconi VC [17]	Code saturation; meaning saturation	Compared 2 approaches to assessing saturation: code and meaning saturation	• Code saturation reached at 9 interviews • Meaning saturation reached at 16-24 interviews
Turner-Bowker DM, Lamoureux RE, Stokes J, et al. [13]	Code saturation	Divided the sample into quartiles and then compared the number of responses elicited from the first 25% of participants to the next 25% of	• 84% of concepts emerged by the 10th interview; 92% emerged by the 15th interview; 97% emerged by the 20th interview; 99% emerged by the 25th

		participants, from the first 50% of participants to the next 25% of participants, and then from the first 75% of participants to the last 25% of participants	interview
Weller SC, Vickers B, Bernard HR, et al. [22]	Thematic saturation; salience	Used a quantitative model to predict the unique number of items contributed by each additional respondent, and saturation was defined as the point where fewer than 1 new item per person would be expected	• The median sample size for reaching saturation was 75 interviews (range = 15-194) • Small samples (n=10) produced 95% of the most salient ideas
Young DS, Casey EA [28]	Code saturation	Used retrospective data from 3 studies; used a random number generator to draw 10 random subsamples of each size from n=5 through n=10 for individual interviews; examined to see what proportion of the codes and larger themes from each original study's full sample were present within each subsample	• Near code saturation reached with a sample size of 6-9 interviews • Partial theme representation required 4-6 interviews • Substantial theme completion required 7-20 interviews

Most studies assessing saturation focused on in-person data collection or did not specify the data collection method. Given recent increases in virtual data collection, studies assessing saturation for virtual interviews are critical to ensure that recommendations regarding sample size are tailored to the mode of data collection.[4] While there is evidence to suggest that the content of data coded from in-person as compared with virtual interviews is conceptually similar,[29] this is a relatively new area of exploration. Rapport may be higher with in-person as compared with virtual interviews,[30] which may impact the amount and type of content generated. Additionally, participants in virtual data collection studies are more geographically diverse and may be more likely to be non-white, less educated, and less healthy than participants in in-person data collection studies[31].

Methods

This study was based on analyses from data collected for 5 FDA-funded studies conducted using virtual platforms, such as Zoom and Adobe Connect, and focused on patients with underlying medical conditions or on healthcare providers who provide primary or specialty care to patients. All platforms used for these interviews offered an audio and video component and allowed for the sharing of stimuli on screen. A brief description of each study is provided in Table 2. Each study's data had been coded and stored using NVivo (version 11, QSR International).

Table 2. Description of studies included in analysis of code saturation: Sample size, eligibility criteria, topics covered, length of interview, number of questions, regions and states covered

Study name	Sample size	General eligibility criteria	Primary objectives	Summary of topics	Length of interview	Number of interview questions	Regions & states covered
Study A	N=30	Patients diagnosed with a condition treated by biologic medications (e.g., cancer, inflammatory bowel disease, diabetes)	Obtain feedback on multimedia educational materials about biosimilar biologic medications.	• Biosimilar awareness • Feedback on educational materials (eg, comprehension, main message, format) • Behavioral intentions	90 minutes	• 37 main questions • Questions identified as high, average, and low priority	Regions: • Northeast • Midwest • South • West States: 14
Study B	N=48	Patients diagnosed with vulvovaginal atrophy or type 2 diabetes	Explore how patients use boxed warnings when making decisions about prescription drugs and how well the warnings meet patients' information needs.	• Prescription drug information needs • Boxed warning awareness, interpretation, and perceptions • Behavioral intentions	30 minutes	13 main questions	Regions: • Northeast • Midwest • South • West States: Not Available
Study C	N=70	Primary care physicians or specialists who write at least 50 prescriptions per week	Assess how primary care physicians and specialists access, understand, and use prescription drug labeling information, including information on labels for drugs that have multiple indications.	• Resources to find information about prescription drugs • Background on prescribing information • Interpretation of language in the prescribing information	60 minutes	36 main questions	Regions: • Northeast • Midwest • South • West States: 26
Study D	N=35	Patients diagnosed with type 2 diabetes	Understand how patients weigh the potential	• Background information on condition • Treatment	60 minutes	20 main questions	Regions: • Northeast • Midwest

			benefits against possible risks and side effects, dosage and administration characteristics, and costs when selecting treatments for chronic health conditions.	decisions and discussion of attributes • Ranking attributes • Condition-specific statements about attributes • Market claims			• South • West States: 9
Study E	N=35	Patients diagnosed with psoriasis	Understand how patients weigh the potential benefits against possible risks and side effects, dosage and administration characteristics, and costs when selecting treatments for chronic health conditions.	• Background information on condition • Treatment decisions and discussion of attributes • Ranking attributes • Condition-specific statements about attributes • Market claims	60 minutes	21 main questions	Regions: • Northeast • Midwest • South • West States: 9

Ethical Considerations

This project was determined to be not research with human subjects by RTI's IRB. The original five studies that this project is based on were reviewed by RTI's IRB and were determined to be exempt under category 2ii. Participants in these studies were provided information about measure used to protect their privacy and the confidentiality of their data in the study's consent forms. All participants were provided compensation for their time (the amount and type varied by study).

Data Preparation and Analysis

We established and applied a systematic approach to analyze all 5 study datasets. Our analytic approach was organized into 2 stages—data preparation and data analysis.

Data Preparation. First, because previous interviews sometimes influence moderator probes—for example, the moderator asks a follow-up question based on something they heard in a previous interview—we sorted interviews from each study by interview order. We then extracted code- and interview-specific data from the NVivo databases—including transcript name, code name, number of files coded, number of associated parent and child codes, and number of coding references—and compiled these data in a Microsoft Excel file. We then updated the Excel file with important code and interview characteristics, including the order in which interviews were conducted, whether each code was directly (i.e., child codes) or indirectly (i.e., parent codes) applied to transcripts (in a tiered coding scheme, direct codes are

those that have no child codes, whereas indirect codes function as “parents” that have additional codes nested beneath them), and the point at which each code was first applied to an interview. Finally, we created pivot tables within each Excel file to compile the data.

Data Analysis. Once the data were compiled, we examined the data summaries to determine when true saturation and near saturation occurred during data collection. True saturation was defined as 100% of all applied codes being used; near saturation was defined as 90% of all applied codes being used. We calculated saturation separately for each study’s dataset, and we calculated saturation separately for all codes (i.e., parent and child codes) as compared with direct codes (i.e., child codes only). We identified true saturation and near saturation points by calculating the cumulative percentage of new codes for each interview, flagging when 100% and 90% of applied codes had been used.

Results

True and Near Saturation

The number of virtual interviews used across the 5 studies ranged from 30 to 70 (Table 3). True saturation (100% use of all applied codes) was reached in the final or near final interview (Figure 1), suggesting that, even with a large sample size, additional interviews are likely to continue uncovering a small number of new codes or findings.

Across all studies, near saturation (90% use of all applied codes) was reached near—and often before—the midpoint of data collection. In other words, only a small number of new codes or findings were uncovered once the first half of the sample had been interviewed. In terms of absolute numbers, the point at which near saturation was reached occurred between 33% and 60% of planned interviews (15-23 interviews). (Table 3). Despite the participants being more geographically, and possibly demographically, diverse compared with typical in-person participants, our findings were similar to previous studies on saturation.[10, 15, 17]

Table 3. Number and percentage of interviews needed to reach true and near saturation by study

		Coding	True saturation		Near saturation	
Study	Total interviews	Total Number of Codes in Codebook	Number of interviews needed	Percentage of interviews needed	Number of interviews needed	Percentage of interviews needed
Study A	30	657	30	100%	18	60%
Study B	48	313	47	98%	21	44%
Study C	70	362	67	96%	23	33%
Study D	35	205	33	94%	15	43%
Study E	35	200	32	91%	15	43%

We examined the types of codes applied after near saturation had been reached. In 4 of the 5 studies, most of these codes (57-62%) represented previously established core concepts or themes, such as a trusted source of information, a behavioral intention, or a recommended change to educational material. Codes representing newly identified concepts (10-15%), other/miscellaneous responses (eg, “in general”) (13-41%), uncertainty or confusion (eg, “don’t know”) (0-11%), or categorization for analysis (eg, correct as compared with incorrect) (0-4%) were less commonly applied after near saturation had been reached.

The overwhelming majority of codes applied after near saturation (73-82%) had already been established in study codebooks before analysis. Only a small number of codes applied after this point (18-27%) were conceptually distinct enough to merit updating the study codebooks by including them. Likewise, most of the codes used after near saturation (44-64%)

were applied to only a single interview. Far fewer codes were applied to 2 interviews (0-27%), three interviews (0-21%), or 4 interviews (0-21%).

Study B was an outlier in terms of codes applied after near saturation. This study had fewer codes representing core established concepts (28%) and more codes representing newly identified concepts (24%) or providing categorization for analysis (10%). The study also had a much higher proportion of conceptually new codes (69%) that were added to the study codebook during analysis. These differences may be because the study sampled 2 populations with very different medical conditions (i.e., type 2 diabetes as compared with vulvovaginal atrophy), leading to a broader range of applied codes.

In examining the relationship between the number of codes in the codebook for each study, the study with the most codes (Study A; 657 codes) required the largest number of interviews to reach both true saturation and near saturation. However, this pattern did not hold true for the remainder of the studies. The study with the next highest number of codes (Study C; 362 codes) was third to reach true saturation and last to reach near saturation.

Parent and Child Codes

All 5 study codebooks included both parent (i.e., top-level codes) and child codes (i.e., subcodes). We examined saturation using 2 analytic lenses—(1) all codes (parent and child) and (2) parent codes only—to determine if there were differences in when saturation was reached. We found no differences in when true saturation was reached. However, near saturation was reached slightly later (i.e., after an additional 3 to 4 interviews) when examining only parent codes (Figure 2).

Differences by Study

Three of the studies had codebooks that consisted almost entirely of deductive (i.e., concept-driven) codes, whereas the codebooks in the remaining 2 studies contained a mix of both deductive and inductive (i.e., data-driven) codes. Although the results were largely consistent across the 5 studies, as expected, the studies that relied heavily on deductive coding reached both true saturation and near saturation sooner. This finding suggests that studies using more inductive coding and analytic techniques may require slightly larger sample sizes to reach saturation.

Structure of Interview Guide

Although all the studies used a semistructured interview guide, the level of structure varied across studies. The 3 studies (i.e., studies C, D, and E) that had a more structured interview guide (eg, questions for which participants were asked their preference among discrete choices or the range of likely answers was limited) reached both true saturation and near saturation sooner. In fact, the study with the most structured guide reached near saturation the soonest, although it fell in the middle for true saturation. This finding suggests that studies using a less structured interview guide may need to conduct more interviews to reach an acceptable level of saturation.

Discussion

Principal Findings

Although true saturation was not reached until the final interview or close to the final interview, near saturation was reached much sooner, ranging from just below to just above the midpoint of data collection, with most of the studies falling just below the midpoint. Although

additional interviews conducted after near saturation may result in new information, our findings suggest there may be diminishing returns relative to the resources expended. We have identified several study characteristics that researchers can consider when making decisions on sample size for virtual interviews.

Although our findings were mostly consistent across the 5 studies we examined, near saturation was reached sooner on the studies that consisted of largely deductive codes compared with those that had a greater number of inductive codes. Consequently, researchers should consider their analytic approach when determining sample size. Studies that intend for the coding scheme to be iterative throughout the coding process may want to err on the side of having a slightly higher sample size than if the codebook is expected to consist largely of deductive codes tied to the interview guide.

These studies ranged in length from 30 minutes to 90 minutes, and a majority (n=3) lasted 60 minutes. Although the 90-minute study reached both true saturation and near saturation at the latest point, the shortest interview (at 30 minutes) required the second highest number of interviews to reach both saturation points. Although length of the interview may be a minor consideration, the level of structure of the interview guide and the types of codes used seem to be larger drivers.

Our findings point to the need for a slightly higher number of interviews to reach an acceptable level of saturation—categorized by us as near code saturation—than what has been found in other studies. For example, Guest et al.[19] found that 6 interviews were enough to get high-level themes, reaching a plateau at 10 to 12 interviews. Similarly, Young and Casey[28] found that near code saturation was reached at 6 to 9 interviews.

Our findings also build on previous studies looking at saturation for in-person data collection conducted at a small number of sites. Data from our studies included participants from all US Census Bureau regions, which provides support that these findings may be more generalizable than previous studies.

Limitations

Our study had several limitations. First, our analysis was conducted on a sample of 5 studies that had similarities. All the studies were related to the medical field, and our study populations (patients with an identified medical condition and healthcare providers) were knowledgeable about the topics discussed. Second, all the studies were conducted using semistructured interview guides that leaned toward being more structured (i.e., interviewers largely stuck to scripted probes as compared with guides that allow for unscripted follow-up probes and unstructured conversations). Additionally, all the studies used a similar approach to coding by using a mix of both deductive and inductive codes (though to varying extents). Consequently, studies with a less structured approach to both the interview and coding process may yield different results. Finally, all our studies are broadly classified as social science research. The findings for other fields of inquiry, such as economic or medical studies, may differ.

Conclusions

Saturation is an important consideration in planning and conducting qualitative research, yet there is no definitive guidance on how to define and measure saturation, particularly for virtual data collection, which allows for data to be collected from a more geographically diverse sample. Our study provides support that near saturation may be a sufficient measure to target and that conducting additional interviews after that point may result in diminishing returns. Factors to consider in determining how many interviews to conduct include the structure and type of questions included in the interview guide, the coding structure, and the population being studied. Studies with less structured interview guides, studies that rely heavily on

inductive coding and analytic techniques, and studies that include populations that may be less knowledgeable about the topics discussed may require a larger sample size to reach an acceptable level of saturation. Rather than trying to reach consensus on the number of interviews needed to achieve saturation in qualitative research overall, we recommend that future research should explore saturation within different types of studies, such as different fields of inquiry, subject matter, and populations being studied. Creating a robust body of knowledge in this area will allow researchers to identify the guidance that best meets the needs of their work.

Acknowledgements

RTI-affiliated authors received support for the development of this manuscript from the RTI Fellow's Program under RTI Fellow, Leila Kahwati, MD, MPH. All studies included in the analyses were funded by the Food and Drug Administration (FDA). We would like to thank the following FDA staff for their contribution to this research: Kit Aikin, Kevin Betts, Amie O'Donoghue, and Helen Sullivan.

Data Availability

The datasets analyzed for this study are available from the corresponding author upon request.

Conflicts of Interest

None declared.

Abbreviations

FDA Food and Drug Administration

References

1. Miles MB, Michael Huberman A, Saldana J. Qualitative data analysis: a methods sourcebook. SAGE Publications.
2. Trochim WMK, Donnelly JP. The research methods knowledge base. 3rd ed. Atomic Dog; 2008.
3. Keen S, Lomeli-Rodriguez M, Joffe H. From challenge to opportunity: virtual qualitative research during COVID-19 and beyond. Int J Qual Methods; Jan-Dec 2022; (21):16094069221105075. doi:10.1177/16094069221105075
4. Roberts JK, Pavlakis AE, Richards MP. It's more complicated than it seems: virtual qualitative research in the COVID-19 era. Int J Qual Methods; 2021;(20):1-20. doi:10.1177/16094069211002959
5. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol; Nov 21 2018;(18)(1):148.

doi:10.1186/s12874-018-0594-7
6. Lobe B, Morgan DL, Hoffman K. A systematic comparison of in-person and video-based online interviewing. Int J Qual Methods; 2022;(21)doi:10.1177/16094069221127068
7. Constantinou CS, Georgiou M, Perdikogianni M. A comparative method for themes saturation (CoMeTS) in qualitative interviews. Qual Res; 2017;(17)(5):571-588. doi:10.1177/1468794116686650
8. Francis JJ, Johnston M, Robertson C, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol Health; 2010;(25)(10):1229-45. doi:10.1080/08870440903194015
9. Galvin R. How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge? Journal of Building Engineering; 2015;(1):2-12. doi:10.1016/j.jobe.2014.12.001
10. Guest G, Namey E, Chen M. A simple method to assess and report thematic saturation in qualitative research. PLoS One; 2020;(15)(5):e0232076. doi:10.1371/journal.pone.0232076
11. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. Soc Sci Med; Jan 2022;(292):114523. doi:10.1016/j.socscimed.2021.114523
12. Sebele-Mpofu FY, Serpa S. Saturation controversy in qualitative research: complexities and underlying assumptions. a literature review. Cogent Soc Sci; 2020;(6)(1)doi:10.1080/23311886.2020.1838706
13. Turner-Bowker DM, Lamoureux RE, Stokes J, et al. Informing a priori sample size estimation in qualitative concept elicitation interview studies for clinical outcome assessment instrument development. Value Health; 2018;(21)(7):839-842. doi:10.1016/j.jval.2017.11.014
14. Coenen M, Stamm TA, Stucki G, Cieza A. Individual interviews and focus groups in patients with rheumatoid arthritis: a comparison of two qualitative methods. Qual Life Res; Mar 2012;(21)(2):359-70. doi:10.1007/s11136-011-9943-2
15. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods; 2016;(18)(1):59-82. doi:10.1177/1525822x05279903
16. Hagaman AK, Wutich A. 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; 2016;(29)(1):23-41. doi:10.1177/1525822x16640447
17. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? Qual Health Res; Mar 2017;(27)(4):591-608. doi:10.1177/1049732316665344
18. Lowe A, Norris AC, Farris AJ, Babbage DR. Quantifying thematic saturation in qualitative data analysis. Field Methods; 2018;(30)(3):191-207. doi:10.1177/1525822x17749386
19. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. Field Methods; 2006;(18)(1):59-82. doi:10.1177/1525822x05279903
20. van Rijnsoever FJ. (I can't get no) saturation: a simulation and guidelines for sample sizes in qualitative research. PLoS One; 2017;(12)(7):e0181689. doi:10.1371/journal.pone.0181689
21. O'Reilly M, Parker N. 'Unsatisfactory saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. Qual Res; 2012;(13)(2):190-197. doi:10.1177/14687941124461
22. Weller SC, Vickers B, Bernard HR, et al. Open-ended interview questions and saturation.

PLoS One; 2018;(13)(6):e0198606. doi:10.1371/journal.pone.0198606
23. Green J, Thorogood N. Chapter 4: In-depth interviews. Qualitative methods for health research. 2nd ed. Sage Publications; 2004:chap 198-202.
24. Fusch P, Ness L. Are we there yet? data saturation in qualitative research. Qual Rep; 2015;(20):1408-1416.
25. Bowen G. Naturalistic inquiry and the saturation concept: a research note. Qual Res; 2008;(8)(1):137–152.
26. Fugard AJB, Potts HWW. Supporting thinking on sample sizes for thematic analyses: a quantitative tool. Int J Soc Res Methodol; 2015;(18)(6):669-684. doi:10.1080/13645579.2015.1005453
27. Sim J, Saunders B, Waterfield J, Kingstone T. Can sample size in qualitative research be determined a priori? Int J Soc Res Methodol; 2018;(21)(5):619-634. doi:10.1080/13645579.2018.1454643
28. Young DS, Casey EA. An examination of the sufficiency of small qualitative samples. Soc Work Res; 2019;(43)(1):53-58.
29. Namey E, Guest G, O'Regan A, Godwin CL, Taylor J, Martinez A. How does mode of qualitative data collection affect data and cost? Findings from a quasi-experimental study. Field Methods; 2019;(32)(1):58-74. doi:10.1177/1525822x19886839
30. Namey E, Guest G, O'Regan A, Godwin CL, Taylor J, Martinez A. How does qualitative data collection modality affect disclosure of sensitive information and participant experience? Findings from a quasi-experimental study. Qual Quant; 2022;(56)(4):2341-2360. doi:10.1007/s11135-021-01217-4
31. Rupert DJ, Poehlman JA, Hayes JJ, Ray SE, Moultrie RR. Virtual versus in-person focus groups: comparison of costs, recruitment, and participant logistics. J Med Internet Res; Mar 22 2017;(19)(3):e80. doi:10.2196/jmir.6980

Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/987ca6a8c01b5979512f22f42cf1799b.docx

Untitled.
URL: http://asset.jmir.pub/assets/c3cb42c4021a8bd6a3c82e18c5f2e13f.docx

Untitled.
URL: http://asset.jmir.pub/assets/3f80ec350528bdd5b269a9c42f627d96.docx

Multimedia Appendixes

Untitled.
URL: http://asset.jmir.pub/assets/cc361ac6d5fe31cd1ab19c25734f95b2.docx