Social Science & Medicine 292 (2022) 114523

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Sample sizes for saturation in qualitative research: A systematic review of empirical tests

Monique Hennink [a,*], Bonnie N. Kaiser [b]

[a] Hubert Department of Global Health, Rollins School of Public Health, Emory University, 1518 Clifton Rd, Atlanta, GA, 30322, USA
[b] Department of Anthropology and Global Health Program, University of California San Diego, 9500 Gilman Drive 0532, La Jolla, CA, 92093, USA

ARTICLE INFO

Keywords:
Sample size
Saturation
Qualitative research
Interviews
Focus group discussions

ABSTRACT

Objective: To review empirical studies that assess saturation in qualitative research in order to identify sample sizes for saturation, strategies used to assess saturation, and guidance we can draw from these studies.
Methods: We conducted a systematic review of four databases to identify studies empirically assessing sample sizes for saturation in qualitative research, supplemented by searching citing articles and reference lists.
Results: We identified 23 articles that used empirical data (n = 17) or statistical modeling (n = 6) to assess saturation. Studies using empirical data reached saturation within a narrow range of interviews (9–17) or focus group discussions (4–8), particularly those with relatively homogenous study populations and narrowly defined objectives. Most studies had a relatively homogenous study population and assessed code saturation; the few outliers (e.g., multi-country research, meta-themes, "code meaning" saturation) needed larger samples for saturation.
Conclusions: Despite varied research topics and approaches to assessing saturation, studies converged on a relatively consistent sample size for saturation for commonly used qualitative research methods. However, these findings apply to certain types of studies (e.g., those with homogenous study populations). These results provide strong empirical guidance on effective sample sizes for qualitative research, which can be used in conjunction with the characteristics of individual studies to estimate an appropriate sample size prior to data collection. This synthesis also provides an important resource for researchers, academic journals, journal reviewers, ethical review boards, and funding agencies to facilitate greater transparency in justifying and reporting sample sizes in qualitative research. Future empirical research is needed to explore how various parameters affect sample sizes for saturation.

1. Introduction

Saturation is the most common guiding principle for assessing the adequacy of purposive samples in qualitative research (Morse, 1995, 2015; Sandelowski, 1995). However, guidance on assessing saturation and the sample sizes needed to reach saturation have been vague. Until recently, saturation had not been empirically assessed with different types of qualitative data. A growing interest in empirical assessment of saturation has now generated a body of research on the topic, making it an opportune time to synthesize it and identify what we can learn from it. This systematic review sought to identify studies that empirically assess saturation in qualitative research, to identify sample sizes needed for saturation, strategies used to assess saturation, and guidance we can draw from these studies.

The concept of saturation was developed by Glaser and Strauss (1967) as '*theoretical saturation*' and was part of their influential grounded theory approach to qualitative research. Grounded theory focuses on developing sociological theory from textual data to explain social phenomena. Within this approach, theoretical saturation refers to "the point at which gathering more data about a theoretical construct reveals no new properties, nor yields any further theoretical insights about the emerging grounded theory" (Bryant and Charmaz, 2007 p.611). Thus, it is the point in data collection when all important issues or insights are exhausted from data, which signifies that the conceptual categories that comprise the theory are 'saturated', so that the emerging theory is comprehensive and well-grounded in data. Theoretical saturation is also embedded in an iterative process of concurrently sampling, collecting data, and analyzing data (Sandelowski, 1995), whereby data

* Corresponding author.
E-mail addresses: mhennink@emory.edu (M. Hennink), bnkaiser@ucsd.edu (B.N. Kaiser).

https://doi.org/10.1016/j.socscimed.2021.114523
Received 19 July 2021; Received in revised form 22 October 2021; Accepted 31 October 2021
Available online 2 November 2021

continuously inform sampling until saturation.

Although most qualitative research does not follow a grounded theory approach, the concept of saturation is widely used in other approaches to qualitative research, where it is typically called '*data saturation*' or '*thematic saturation*' (Hennink et al., 2017). This broader application of saturation focuses more on assessing sample size rather than the adequacy of data to develop theory (as in theoretical saturation). When used in the broader context, saturation refers to the point in data collection when no additional issues or insights are identified and data begin to repeat so that further data collection is redundant, signifying that an adequate sample size is reached. Saturation is an important indicator that a sample is adequate for the phenomenon studied – that data collected have captured the diversity, depth, and nuances of the issues studied – and thereby demonstrates content validity (Francis et al., 2010). Reaching saturation has become a critical component of qualitative research that helps make data collection robust and valid (O'Reilly and Parker, 2013). Moreover, saturation is "the most frequently touted guarantee of qualitative rigor offered by authors to reviewers and readers" (Morse, 2015, p. 587). In this review, we focus on saturation in the broader context, since less is known about adequate sample sizes for saturation when used outside of the parameters of grounded theory.

Despite the importance of saturation to support the rigor of qualitative samples, there is a consistent lack of transparency in how sample sizes are justified in published qualitative research (Morse, 1995; Guest et al., 2006; Kerr et al., 2010; Carlsen and Glenton, 2011; Hennink et al., 2017). Although saturation is the most commonly cited justification for an adequate sample size (Morse, 1995, 2015), details of how saturation was assessed and the grounds on which it was determined are largely absent in qualitative studies. Vasileiou et al. (2018) conducted a systematic review of qualitative studies using in-depth interviews in health-related journals over a 15-year period and found the vast majority of articles provided no justification for their sample size. Where justifications were given, saturation was cited in 55% of articles; however, claims of saturation were "never substantiated in relation to procedures conducted in the study itself" (p. 12); only further citations of other literature were given that moved away from the study at hand. Similarly, a systematic review of 220 studies using focus group discussions (Carlsen and Glenton, 2011) found that 83% used saturation to justify their sample size; however, they provided only superficial reporting of how it was achieved, including unsubstantiated claims of saturation and references to achieving saturation while still using a predetermined sample size. Another study (Francis et al., 2010) reviewed articles in the journal *Social Science and Medicine* over 16 months and found most articles claimed they had reached saturation but provided no clarity on how saturation was defined, achieved, or justified. Marshall et al. (2013) also reviewed 83 qualitative studies and found saturation was not explained in any study. There are increasing concerns over claims of saturation without study-based explanations of how it was assessed or determined. Unsubstantiated claims of reaching saturation undermine the value of the concept. In part, this lack of transparency may reflect the absence of published guidance on how to assess saturation (Morse, 1995; Guest et al., 2006). In this review, we seek to identify the strategies used to assess saturation in empirical research, which may encourage greater transparency in reporting saturation in qualitative studies.

In addition, guidance on specific sample sizes needed to reach saturation in different qualitative methods has been absent or vague in the methodological literature, providing only general "rules of thumb" that are rarely evidence-based (Morse, 1995; Guest et al., 2006; Kerr et al., 2010; Bryman, 2012; Hennink et al., 2019). As research empirically assessing saturation begins to fill this gap, it allows us to provide much-needed empirically based guidance on sample sizes for saturation in qualitative research.

In this systematic review, we aim to synthesize empirical studies that assess saturation in qualitative data. In particular, we aim to document strategies used to assess saturation, identify sample sizes needed to reach saturation using different qualitative methods, and suggest guidance on sample sizes for qualitative research. To our knowledge, this is the first systematic review on empirical studies of saturation and therefore provides a valuable resource for researchers, academic journals, journal reviewers, ethical review boards, and funding agencies that review qualitative research. Researchers can refer to our results when estimating an appropriate sample size in research proposals and protocols, which may lead to more efficient use of research resources and clearer justifications for proposed sample sizes. Similarly, our results may provide evidence-based expectations regarding adequate sample sizes for qualitative research to guide those who review and fund research.

2. Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines in conducting and reporting our systematic review (Moher et al., 2009). Fig. 1 shows the number of articles identified, screened, and included. We used a two-stage search process, including database searches and citation searches.

First, we used four databases – PubMed, Embase, Sociological Abstracts, and CINAHL – to search for articles or book chapters that included "saturation" and one of the following terms in the title, abstract, or key words/index: "interview," "focus group," "qualitative," or "thematic" (see Supplemental Table for full search terms). Search results were limited to English-language and human studies. Database searches were conducted on January 31 – February 1, 2019 and updated July 10, 2020. Both authors independently screened all article titles, abstracts, and, where needed, full texts to determine eligibility. Discrepancies were discussed and resolved by consensus. To be eligible for inclusion, studies needed to: a) use empirical data to assess saturation in qualitative research or use a statistical model to determine saturation using hypothetical data, b) focus on saturation outside of grounded theory, c) be published in journal articles or book chapters, and d) be available in English. Sixteen articles were included from database searches.

Second, we conducted citation searches by reviewing the reference lists of included articles and using the "cited by" search option in Google Scholar to identify further records meeting the inclusion criteria. For studies with more than 250 citing articles on Google Scholar, we searched within citing articles for "saturation" and reviewed the first 250 results (which are ordered by relevance). An additional seven articles were included during this step.

We extracted the following information from the 23 eligible articles: a) meta-data about the article (author, journal, year), b) information about data used (hypothetical vs. empirical; interviews, focus group discussions, etc.), research objective, sample size, study population (homogenous, heterogenous), and whether data collection was iterative, c) information about saturation, including: definition, goal, data randomization, strategy to assess saturation, sample size for saturation, and level of saturation achieved (e.g., 90% of codes), and d) additional information (limitations, any parameters of saturation suggested). Both authors independently extracted data from 6 articles and discussed results. This was done to identify any issues with the data extraction categories, such as lack of clarity or redundancy, as well as to establish reliability between the two authors. Each remaining article then underwent data extraction by one of the two authors.

We analyzed results separately for studies using empirical data to assess saturation versus those using statistical models. We analyzed sample sizes for saturation by qualitative method: interviews or focus group discussions. We conducted comparisons of saturation by homogeneity of the study population and randomization of data to identify any patterns.

3. Results

Our systematic review identified 23 articles assessing saturation for

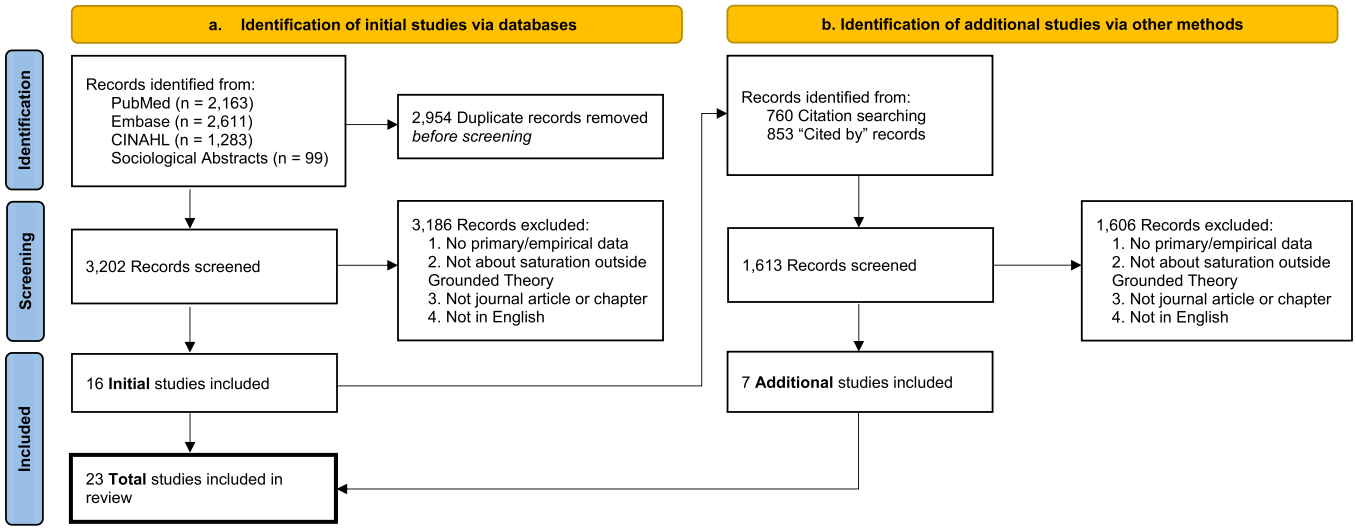

Fig. 1. PRISMA Flow Diagram of Systematic Review Search Procedures.

qualitative research. All articles were published between 2006 and 2020, with the majority (87%, 20/23) published since 2014. Many articles were published in research methodology journals (43%, 10/23) and others in social science (6/23) or topical journals (7/23) (e.g., engineering, computing, natural resources). We categorized the articles into those assessing saturation using empirical data (Table 1, 17 articles) and those using statistical modeling to predict saturation (Table 2, 6 articles). Since these approaches and results are not comparable, we report each separately below.

3.1. Approaches to assessing saturation

3.1.1. Empirically based tests

Table 1 summarizes 17 articles that assess saturation using empirical data. Some articles used multiple datasets to assess saturation and report the results of each separately; therefore, Table 1 shows 23 tests from 17 articles (NB: while these studies were not conducting experimental tests, we use the term 'test' for brevity to refer to individual studies using empirical data, as opposed to statistical modeling, to assess saturation). Most articles used data from in-depth interviews (10/17) or focus group discussions (4/17); two articles used both types of data, and one article (Weller et al., 2018) used free list data. We excluded the article by Weller et al. in our analysis because free list data are not comparable to free-flowing narrative data from interviews and focus group discussions. We therefore use the denominator of 16 when describing all articles and 22 when describing the datasets and results of all tests with empirical data.

The original research objective for each dataset used in the tests varied, but most studies (14/16) focused on health issues, such as experiences of a specific health condition (e.g., sickle cell disease, multiple sclerosis, Paget's disease), health service, or intervention (e.g., genetic screening, violence prevention, lifestyle interventions, patient retention). These research objectives are typical of much qualitative heath research. The sample size of the datasets used varied from 14 to 132 interviews and 1 to 40 focus groups. All datasets except one (Francis et al., 2010) had a sample that was much larger than the sample ultimately needed for saturation, making them effective for assessing saturation. Francis et al. (2010) report saturation was reached at exactly the sample size of the study for both datasets used. Most datasets (18/22) had a homogenous study population, such as patients with a specific disease (e.g., HIV, rheumatoid arthritis, sickle cell) or from a specific demographic group (e.g., male nurses, medical students, South Asian adults, African American men). The remaining datasets had more heterogeneous samples, such as men aged 20–72 across the US or youths aged 14–18.

Authors described their goal of saturation in two ways, either as saturation of individual codes or categories. Although terminology varied across articles, codes were typically described as individual issues, topics, or items in data, while categories represented higher-order groupings of issues (e.g., broader themes, meta-themes, concepts). Forty-four percent (7/16) of articles sought saturation of codes, 31% (5/16) saturation of categories, and 25% stated both.

Where saturation was defined, authors used similar definitions. Overall, saturation was described as the point at which little or no relevant new codes and/or categories were found in data, when issues begin to be repeated with no further understanding or contribution to the study phenomenon, its dimensions, nuances, or variability. Some articles further specified that saturation should be confirmed only after no new issues were found in two or three consecutive interviews or focus groups (Coenen et al., 2012; Francis et al., 2010; Morse et al., 2014) or that it was determined by two researchers (Morse et al., 2014). Over half of articles (56%, 9/16) randomized the order of data for analysis to account for interview order, which might influence saturation. Some compared saturation between the randomized order of interviews and the actual order in which interviews were conducted, while others calculated saturation across multiple randomized orderings of data to identify an average.

Various strategies were used to assess saturation. These are categorized in Table 1 and the categories described in Table 3. Most articles (75%, 12/16) used a single strategy to assess saturation. All articles used some form of code frequency counts to assess saturation (including code frequency counts, comparative method, stopping criterion, higher-order groupings), and four articles used another approach in addition to code frequency counts and compared saturation for each (Hennink et al., 2017, 2019; Constantinou et al., 2017; Hagaman and Wutich, 2017). Many articles (37% 6/16) used only code frequency counts to assess saturation, which involved counting codes in successive transcripts or sets of transcripts until the frequency of new codes diminishes, signaling saturation is reached. Three articles (18%, 3/16) added specific additional elements to code frequency counts, such as batch comparisons, a stopping criterion, or counting higher-order groupings of codes, such as meta-themes or categories of codes rather than individual codes. Three articles (Hennink et al., 2017, 2019; Nascimento et al., 2018) used 'code meaning' to assess saturation, an entirely different approach from code frequency counts. This approach focused on reaching a full *understanding* of issues in data as the indicator that saturation is reached, by assessing whether the issue, its dimensions, and nuances are fully identified and understood. Two articles (Hennink et al., 2017, 2019) then compared

Table 1
Articles assessing saturation using empirical data.

ARTICLE	DATA USED			SATURATION RESULTS			
Author, Date, Journal	Research Objective	Sample Size	Study Population	Saturation Goal[b]	Data Randomized	Strategy to Assess Saturation[c]	Sample Size for Saturation
Type of Data Used: In-Depth Interviews							
Ando et al. (2014) *Comprehensive Psychology*	Influences on quality of life for people with neurological conditions	39	Homogenous	Codes & Categories	No	Code Freq. Counts	12 interviews for 92% of codes
[a] Coenen et al. (2012) *Qual. Life Research*	Daily life challenges for patients with rheumatoid arthritis	21	Homogenous	Categories	No	Higher Order Groupings	9 interviews (inductive) 12 interviews (deductive)
Constantinou et al. (2017) *Qualitative Research*	Medical students' beliefs and attitudes towards psychotherapy	12	Homogenous	Categories	Yes	Code Freq. Counts & Higher Order Groupings	5 interviews (consecutive) 8 interviews (random)
[a] Francis et al. (2010) *Psychology and Health*	How do doctors manage sore throat without antibiotics	14	Homogenous	Categories	No	Stopping Criterion	14 interviews
[a] Francis et al. (2010) *Psychology and Health*	Acceptability of genetic screening for relatives of patients with Padgett's disease	17	Homogenous	Categories	No	Stopping Criterion	17 interviews
Guest et al. (2006) *Field Methods*	Perceptions of social desirability bias in self-reported reproductive health behavior	60 (2 countries)	Homogenous	Codes	No	Code Freq. Counts	12 interviews for 88% of codes
[a] Guest et al. (2020) *PLOS ONE*	Health seeking behaviors of African American men in southeast US	40	Homogenous	Codes	Yes	Stopping Criterion	Depends on parameters 11-14 interviews for ~90% of themes at 0% threshold
[a] Guest et al. (2020) *PLOS ONE*	Medical risks during pregnancy amongst mothers in southeast US	48	Homogenous	Codes	Yes	Stopping Criterion	Depends on parameters 11-14 interviews for ~90% of themes at 0% threshold
[a] Guest et al. (2020) *PLOS ONE*	Women at high risk of HIV in Kenya and South Africa	60	Heterogeneous	Codes	Yes	Stopping Criterion	16 interviews for ~80% of themes
Hagaman and Wutich (2017) *Field Methods*	Cultural understandings of water problems and solutions	132 (4 countries)	Heterogenous	Codes & Categories	Yes	Higher Order Groupings & Stopping Criterion	16 interviews (top 3 themes within country); 20–40 interviews (meta-themes across countries)
Hennink et al. (2017) *Qualitative Health Research*	Influences on patient retention in HIV care	25	Homogenous	Codes	Yes	Code Freq. Counts & Code Meaning	9 interviews for 91% of codes 16-24 interviews for meaning saturation
Namey et al. (2016) *Am. J. Evaluation*	Health seeking behaviors of African American men in Durham, NC	40	Homogenous	Codes	Yes	Code Freq. Counts	16 interviews for 90% of codes
Nascimento et al. (2018) *Revista Brasileira de Enfermagem*	Daily functions of school children with sickle cell disease	15	Homogenous	Categories	No	Code Meaning	11 interviews
Turner-Bowker et al. (2018) *Value in Health*	Patient experiences of symptoms in acute or chronic health conditions	26	Homogenous	Categories	No	Comparative Method	15 interviews for 92% of concepts
[a] Young and Casey (2019) *Social Work Research*	Challenges and strategies for engaging men in prevention of gender-based violence	27	Heterogenous	Codes & Categories	Yes	Code Freq. Counts	9 interviews for 90% of codes
[a] Young and Casey (2019) *Social Work Research*	Social workers perspectives of success in working with the US justice system	15	Homogenous	Codes & Categories	Yes	Code Freq. Counts	9 interviews for ~90% of codes
Type of Data Used: Focus Group Discussions							
[a] Coenen et al. (2012) *Qual. Life Research*	Daily life challenges for patients with rheumatoid arthritis	NS	Homogenous	Categories	No	Higher Order Groupings	5 groups (inductive & deductive)
Guest et al. (2016) *Field Methods*	Health seeking behaviors of African American men in Durham, NC	40 groups	Homogenous	Codes	Yes	Code Freq. Counts	4.3 groups (mean) or 3–6 groups for 90% of codes
Hancock et al. (2016) *The Qualitative Report*	Experiences of male registered nurses seeking employment	1 group (asynchronous)	Homogenous	Codes & Categories	No	Stopping Criterion	1 asynchronous online group
Hennink et al. (2019) *Qualitative Health Research*	Design a lifestyle intervention for diabetes in South Asian Americans	10 groups	Homogenous	Codes	Yes	Code Freq. Counts & Code Meaning	4 groups for 94% of codes 2 per strata for meaning saturation
Morse et al. (2014) *Society & Natural Resources*	Important places for recreation/livelihoods	19 groups	Heterogenous	Codes	Yes	Code Freq. Counts	16 groups for 90% saturation at all hotspots

(*continued on next page*)

Table 1 (*continued*)

ARTICLE	DATA USED			SATURATION RESULTS			
Author, Date, Journal	Research Objective	Sample Size	Study Population	Saturation Goal[b]	Data Randomized	Strategy to Assess Saturation[c]	Sample Size for Saturation
[a] Young and Casey (2019) *Social Work Research*	Influences on adolescent bystander's responses to dating violence and bullying	12 groups	Heterogenous	Codes & Categories	Yes	Code Freq. Counts	7 groups for complete themes
Type of Data Used: Free Lists							
Weller et al. (2018) *PLoS ONE*	Free listing of items in topical domains	Free-Lists in 28 datasets	Varies by study	Items	No	Most Salient Items	10 interviews for 95% of salient items (named by 20% of participants)

Note: where different levels of saturation are identified, saturation closest to 90% is presented here. Where percentage saturation is not specified, this is due to authors not indicating this or using another measure to determine saturation (e.g. stopping criterion, specific number of repetitions of a code).
NS - not stated.
[a] These studies report results of multiple data sets separately and are included in separate rows.
[b] Codes refers to single-level issues in data; categories refers to any higher order groupings of codes.
[c] See Table 2 for description of these categories.

Table 2
Articles estimating saturation through statistical modeling.

	Data Application	Strategy to Assess Saturation	Parameters and Assumptions	Suggested formula for saturation
Fofana et al. (2020) *PLOS ONE*	Statistical model tested on empirical dataset of interviews (n = 12)	Uses set theory and partial least squares regression to estimate saturation	X_j is the vector of the number of times each theme is coded in the j-th interview B_{PLS} is the vector of regression coefficients E is the matrix of residuals	$(X_{j+1} \dots X_n) = (X_1 \dots X_j)$ $B_{PLS} + E$
Fugard and Potts (2015) *Int. J. Soc. Res. Methodology*	Hypothetical model based on interviews but not tested on empirical data	Uses negative binomial probability distribution to estimate sample needed to reach a certain power (eg, 80% probability to identify a theme) based on several parameters	Assumes random sample. Estimates sample size based on population theme prevalence (known probability of issue/theme in the population of interest) of least prevalent theme, desired number of instances in the data, and desired power.	Various outcomes are provided based on a range of values for model inputs
Galvin (2015) *J. Building Engineering*	Hypothetical model based on interviews but not tested on empirical data	Uses binomial distribution to answer 5 research questions; the most relevant is RQ3: How many interviews to have 95% probability of theme emerging?	Assumes random sample P = probability theme arising in interview R = proportion of theme in population n = # interviews	$n = \frac{ln(1-P)}{ln(1-R)}$
Lowe et al. (2018) *Field Methods*	Statistical model tested on empirical datasets including literature surveys (n = 25), focus groups (n = 3), and interviews (n = 11)	Develops saturation index using generalized estimating equations	R = prevalence of a theme in population P = particular saturation n = # observations Accounts for statistical dependency between observations and likelihood of researcher identifying theme. Assumes order of observations does not influence themes identified. Assumes random sample	$n = \frac{P(R-1)}{R(P-1)}$
Rowlands et al. (2016) *J. Computer Inf. Systems*	Statistical model tested on empirical data of interviews (3 studies: n = 30, 30, 24)	Calculate thematic saturation using lognormal distribution with chosen confidence level	Based on concept analysis using Leximancer program. $\bar{X}^*$ is the geometric mean from the lognormal fit s* is the multiplicative standard deviation from the lognormal fit	For 95% confidence, lognormal expression $= \bar{x}^*(s^*)^2$
Van Rijnsoever (2017) *PLOS ONE*	Hypothetical model based on various data types (e.g., interviews, focus groups, documents) but not tested on empirical data	Uses simulations based on lognormal distribution and 11 parameters	Accounts for random and purposive samples, as well as minimal and maximal information from observations.	Various outcomes are provided based on a range of values for model inputs

saturation using this approach with the code frequency approach.

3.1.2. Statistical models

Table 2 summarizes six articles that used statistical modeling to estimate saturation. These articles used a different approach than those summarized above: they developed a formula to estimate the sample size needed for saturation, which may be used prior to data collection to inform study design. Several formulas were based on similar parameters, such as prevalence of a theme in a population or the desired instances of a theme in data (Fugard and Potts, 2015; Galvin, 2015; Lowe et al., 2018), while others used a lognormal distribution (Rowlands et al., 2016; Van Rijnsoever, 2017) or set theory (Fofana et al., 2020). Many of these studies assumed a random sample, while one accounts for both random and purposive samples (Van Rijnsoever, 2017). Most formulas were developed for interview data, while two articles discussed estimating saturation for various forms of data, including interviews, focus groups, documents, and literature surveys. Half of the formulas were then applied to empirical datasets.

3.2. Sample size for saturation

Fig. 2 shows sample sizes for saturation from empirically based tests using in-depth interview data. The results for each dataset used in the tests (n = 16) are shown as separate data points. Where results are reported at different sample sizes, this is depicted with a line from the lowest to highest sample size reported, and the parameters influencing this range are noted. Where authors report different levels of saturation, saturation closest to 90% is shown for comparability across studies.

Table 3
Strategies to assess saturation in empirical tests.

Type of Approach	Description of Approach	Articles Reporting Strategy
Code Frequency Counts	This approach involves reviewing each interview or focus group transcript and counting the number of new codes in each successive transcript or set of transcripts, until the frequency of new codes diminishes with few or no more codes identified. Several articles additionally randomized the order of data to assess the influence of sequential bias on saturation. Some articles added additional elements to the code frequency counts, such as batch comparison, a stopping criterion or saturation of higher order groupings of data, as outlined below.	Ando et al. (2014); Guest et al. (2006, 2016); Morse et al. (2014); Namey et al. (2016); Hennink et al. (2017 [a], 2019) [a]; Constantinou et al., 2017 [a]); Young and Casey (2019)
Comparative Method	This approach adds a more structured comparison to the code frequency count approach above. It involves reviewing data in pre-determined batches, such as quartiles of data (instead of reviewing each interview separately) and listing all new codes in a saturation table for each batch of data. The subsequent quartile of data is then reviewed and compared to the first quartile to determine any new codes, this comparison of data batches continues until few or no new codes are identified, whereby saturation is achieved.	Turner-Bowker et al. (2018)
Stopping Criterion	This approach adds a stopping criterion to the code frequency count approach above. It involves reviewing an initial sample of interviews (e.g. 6 interviews) or focus groups to identify new codes, and using a pre-determined stopping criterion, which is usually the number of consecutive interviews/groups after the initial sample where no new codes are identified in the sample (e.g. 2 or 3 interviews with no new codes). Saturation is reached when no new codes are identified after the stopping criterion of x interviews after the initial sample, or the number of new codes is under a predetermined threshold (e.g. <5%). In other studies, the stopping criterion was based on repetitions of a code, such as 3 or 5 instances of a particular code or theme were identified.	Francis et al. (2010); Guest et al. (2020); Hagaman and Wutich (2017) [a]; Hancock et al. (2016)
High-Order Groupings	This approach uses a higher order grouping of codes in the code frequency count approach above. It involves counting higher-order groupings of codes such as meta-themes, salient themes or categories. For example, Coenen et al. (2012) counted conceptual categories. Hagaman and Wutich (2017) counted codes to determine the *most prevalent* codes in the data set, then randomized the interview order via bootstrapping to determine the average number of interviews needed to identify the most prevalent codes in data. Weller et al. (2018) focused on identifying saturation for the *most salient* items in data.	Coenen et al. (2012); Hagaman and Wutich (2017) [a]); Constantinou et al. (2017) [a]
Code Meaning	This approach does not focus on counting codes as the basis for determining saturation (as used in the approaches above), instead achieving a full understanding of codes is the indicator of saturation. It involves reviewing an interview and noting each issue (or code) identified, then in subsequent interviews identifying whether any new aspects, dimensions, or nuances of that code are identified, until nothing new is identified and the code has reached saturation. Codes may reach saturation at a different point in the data set.	Hennink et al. (2017·[a], 2019) [a]; Nascimento et al. (2018);

[a] These articles used multiple approaches and are therefore listed twice.

Results show that across 16 tests using various approaches to saturation, the sample size for saturation ranges between 5 and 24 interviews. The lowest sample size for saturation was 5 interviews (Constantinou et al., 2017), in a study with a homogenous study population that was intended to support survey findings and where saturation was sought in broad categories. Together, these study characteristics may explain reaching saturation at 5 interviews. The highest sample sizes for saturation were 20–40 (Hagaman and Wutich, 2017), where saturation of meta-themes across four countries was sought, and 24 (Hennink et al., 2017), where saturation was sought in the meaning of codes, including codes less central to the research question. These saturation goals require more data, which may support the higher sample sizes found for saturation. Excluding these outliers, most datasets reached saturation between 9 and 17 interviews, with a mean of 12–13 interviews, despite using different approaches to assess saturation. Most of these studies had a relatively homogenous study population and varied in their saturation goal of codes, categories, or a combination. Only three studies used a heterogeneous sample. Two of these studies reached saturation at a larger sample size than the mean (at 16 interviews), and one reached saturation at a smaller sample size (at 9 interviews). Therefore, we found no pattern in saturation by this characteristic. Similarly, it was difficult to identify any pattern of saturation by the order of data, since most tests did not compare saturation when analyzing data in the actual interview order with the randomized order. Those that did make a comparison found no difference or a slightly higher sample size for saturation in the random versus actual order of interviews. Both studies that used randomization and those that did not cover the full spectrum of sample sizes seen in our review.

Fig. 3 shows the sample size for saturation from six empirical tests using data from focus group discussions. For comparability, where various levels of saturation are reported, those closest to 90% are shown in the figure. Across all six tests, saturation was reached between 1 and 16 focus groups. Two tests are outliers and thus not comparable to others. At the lower end, Hancock et al. (2016) report on saturation in a single asynchronous, online focus group, and saturation is reported by day and participant. At the higher end, Morse et al. (2014) report reaching saturation at 16 groups; however, they focus on spatial locations rather than codes or themes, which may account for the higher sample size for saturation. The remaining four tests used similar definitions of saturation and reached saturation between 4 and 8 focus groups, with a mean of 5–6 groups. Most tests (4/6) had a homogenous study population but varied in their approach to assessing saturation and the saturation goal of codes or categories. In the two tests using heterogeneous samples, both reached saturation at sample sizes above the mean number of groups (at 7 and 17 groups).

In studies that developed statistical models for saturation that were applied to empirical data, the sample sizes for saturation were similar to those above (Table 2). For example, Rowlands et al. (2016) used the lognormal distribution to estimate saturation in three datasets of interviews, and results found the sample sizes for saturation at 95% confidence to be 10, 10, and 13 interviews. Fofana et al. (2020) used set theory and partial least squares regression to estimate saturation at 12

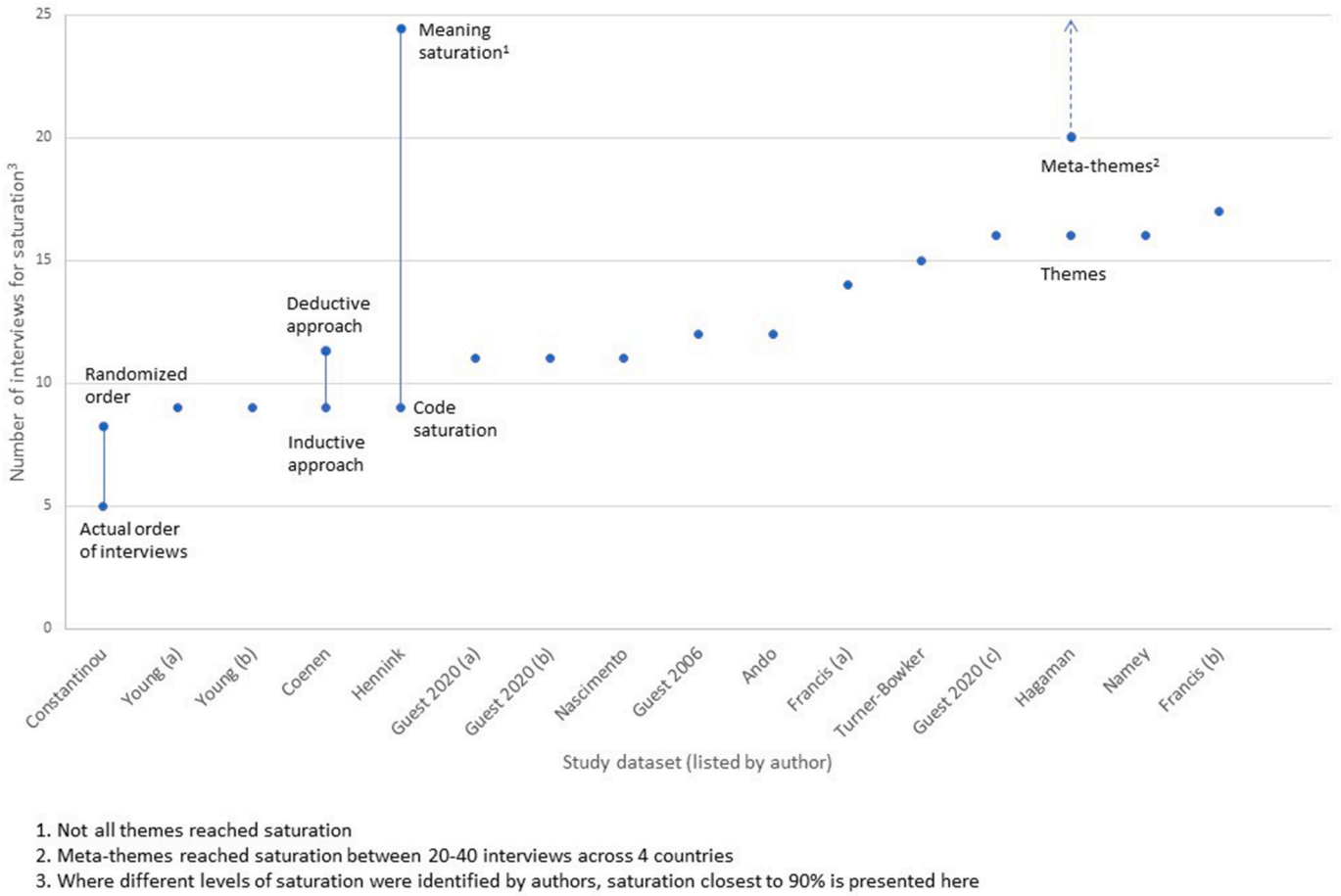

Fig. 2. Sample size for saturation in empirical tests with interview data.

interviews when applied to an empirical dataset.

4. Discussion

This systematic review sought to identify empirical studies that assess saturation, to identify sample sizes needed for saturation, strategies used to assess saturation, and guidelines we can draw from these studies. We identified 23 studies that empirically assessed saturation, with 80% published since 2014. We identified two different approaches to assess saturation: studies that used empirical data and those that used statistical models.

One approach to assessing saturation focused on developing statistical models to estimate sample sizes for saturation prior to data collection. While we applaud efforts to estimate saturation *a priori,* many of the formulas developed are based on implicit assumptions that do not align with the conduct of qualitative research, thereby significantly limiting their utility. Many of these studies use probability-based assumptions, such as having a random sample and knowing the prevalence of a theme in the broader population or the desired instances of a theme in data. Moreover, researchers often do not know these parameters prior to conducting a study, nor is prevalence of items an important focus of qualitative research. Since a statistical formula may be seen as akin to a power calculation familiar to quantitative researchers, we feel that this may provide a misleading veil of scientific authenticity that ultimately cannot be achieved given the misalignment of assumptions with qualitative research. Given our concerns about these approaches, we do not consider them further.

A second approach to assess saturation used empirical data. In all 16 tests of saturation with data from in-depth interviews, saturation was reached in under 25 interviews, more specifically between 9 and 17 interviews excluding outliers. Despite using different approaches to assess saturation, different datasets, varying saturation goals (codes vs categories), and homogenous and heterogeneous study populations, studies still reached saturation within a narrow range of interviews. This demonstrates strong external reliability across the different approaches. Across all tests, an average of 12–13 interviews reached saturation, which is remarkably similar to findings from Guest et al. (2006), one of the first studies to empirically assess saturation, which reported saturation at 12 interviews. We found no clear pattern in saturation by study characteristics, such as homogeneity of the study population, use of randomization, or saturation goal, largely because few studies actually assessed these parameters in their approach. In six tests using data from focus group discussions, saturation was reached by 4–8 groups, a similarly narrow range. Studies using demographic stratification, heterogeneous samples, and broader saturation goals (e.g., code meaning, all themes vs main themes) needed more groups to reach saturation. However, we are cautious about drawing conclusions regarding the influences of these characteristics without more studies with focus group data to compare. Overall, these findings provide much-needed empirical evidence of sample sizes for saturation for different qualitative methods. Despite convergence of saturation within a specific range of interviews or focus groups, we caution not to use these findings as *generic* sample sizes for *any* qualitative study using these methods, or to justify poorly designed or executed qualitative studies, as we discuss below. Instead, we recommend using these results as guidance to consider alongside the specific study characteristics when estimating the sample size for a qualitative study.

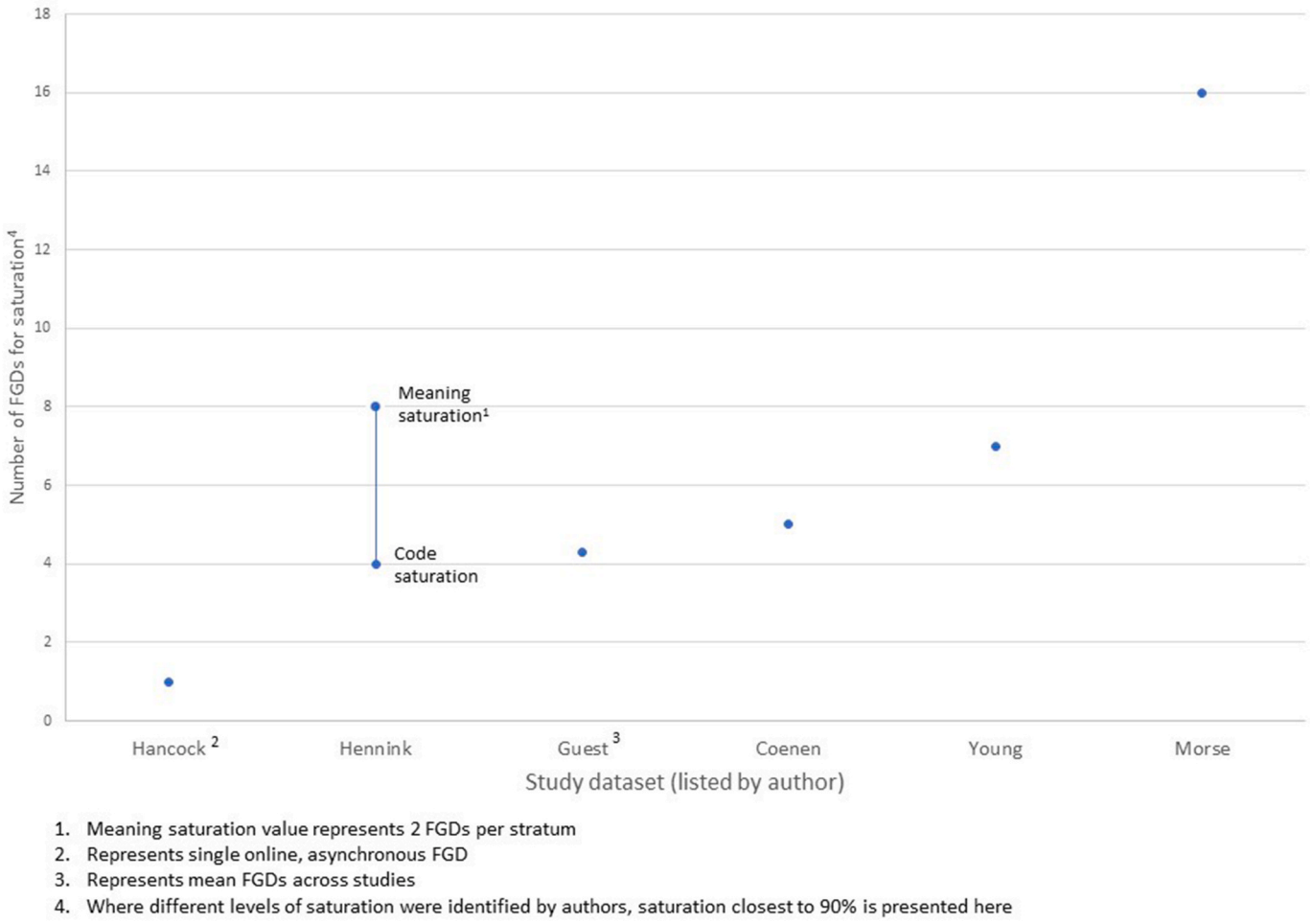

Fig. 3. Sample size for saturation in empirical tests with focus group discussion data.

4.1. Implications for research

The results of our systematic review have several important implications. We focus here only on implications of empirically based studies. These results provide empirical guidance regarding adequate samples sizes for saturation when using interviews and focus group discussions, which can be useful when developing qualitative research proposals. The majority of empirically based studies in our review had a homogenous study population and focused research objectives, so these results cannot be confidently extrapolated to studies using different types of samples or broader goals. Therefore, we recommend using these results as a starting point to identify a potential range of interviews or focus groups then refining the sample size by considering the study characteristics (e.g., study goals, nature and complexity of phenomenon studied, instrument structure, sampling strategy, stratification of sample, researcher's experience in qualitative research, saturation goal, and degree of saturation sought) (Baker and Edwards, 2012; Galvin, 2015; Morse, 1995; see Hennink et al., 2017 for fuller discussion on using study parameters to estimate saturation). These considerations will not only lead to a more tailored sample size for each particular study but also provide clearer justification for the proposed sample size, thereby adding rigor.

Our results also provide researchers with strong empirical evidence to refute the common critique that qualitative sample sizes are 'too small', implying that they are ineffective, although no evidence is usually given for these claims. Our results can be used to demonstrate that 'small' sample sizes *are* effective for qualitative research and to show *why* they are effective – because they are able to reach saturation, the long-held benchmark for an adequate sample size in qualitative research. Furthermore, our results show what a 'small' sample actually is, by providing a range of sample sizes for saturation in different qualitative methods (e.g., 9–17 interviews or 4–8 focus groups). This is important because general advice on sample sizes for qualitative research usually suggest higher sample sizes than this. Reviews of textbooks on qualitative research methodology found that sample size recommendations vary widely, for example 5–60 interviews (Guest et al., 2006; Constantinou et al., 2017; Hagaman and Wutich, 2017) and 2 to 40 focus groups (Guest et al., 2016). More importantly, none of these recommendations is empirically based. Providing evidence-based sample size recommendations, with appropriate caveats, is important. Qualitative samples that are larger than needed raise ethical issues, such as wasting research funds, overburdening study participants, and leading to wasted data (Carlsen and Glenton, 2011; Francis et al., 2010), while samples that are too small to reach saturation reduce the validity of study findings (Hennink et al., 2017). Our results thus provide empirically based samples sizes for saturation that could be included as part of the guidelines in instructional textbooks on qualitative research.

Furthermore, Vasileiou et al. (2018) found that even some qualitative researchers characterized their own sample size as 'small', but this was "construed as a limitation couched in a discourse of regret or apology" (p. 12). Although these authors may be writing to the concerns of more positivist-oriented readers, few defended their 'small' sample on qualitative grounds. We encourage researchers to reflect on our results to more confidently justify their sample sizes using the principles of qualitative research rather than responding to the (mostly inappropriate) concerns of a more dominant positivist paradigm and their

numerical expectations. Sample sizes in qualitative research are guided by data adequacy, so an effective sample size is less about numbers (n's) and more about the ability of data to provide a rich and nuanced account of the phenomenon studied. Ultimately, determining and justifying sample sizes for qualitative research cannot be detached from the study characteristics that influence saturation. Our results echo others, that "rigorously collected qualitative data from small samples can substantially represent the full dimensionality of people's experiences" (Young and Casey, 2019, p.12) and therefore should not be viewed or presented as a limitation when evaluating the rigor of qualitative research.

Our results also provide empirical guidance on effective sample sizes for saturation for reviewers of qualitative research. This may help to refocus the routine practice of criticizing qualitative research for 'small' sample sizes so that reviewers may instead ask researchers to provide more explicit justifications for their sample size by asking, for example: "*why do you have a sample of 40 interviews, when saturation can typically be reached in less than 25 with a homogenous study population such as yours*?" Although, we generally do not support using only numerical guidance in determining an effective sample size for qualitative research, these types of questions reflect a more informed critique that uses available empirical evidence on saturation to challenge researchers to be more transparent in justifying their sample sizes and using the characteristics of each individual study to do so. We therefore encourage qualitative researchers to provide fuller justifications of their sample sizes and urge reviewers of qualitative studies to apply these findings to provide more effective critiques of sample sizes for qualitative research. This may improve the quality of reporting and critiquing qualitative research and move away from often unsubstantiated critiques of 'small' sample sizes.

Our results also synthesize five distinct approaches to assess saturation, including several variations of code frequency counts and assessing code meaning. Qualitative researchers now have an array of strategies to assess saturation during data collection. Numerous reviews of qualitative studies have found that saturation is often used to justify a sample size, but there was an overwhelming lack of transparency in how it was assessed or determined (Carlsen and Glenton, 2011; Francis et al., 2010; Marshall et al., 2013; Vasileiou et al., 2018). This lack of transparency is concerning, particularly given that saturation is hailed as an indicator of quality in qualitative research. It suggests that saturation is being used as a "mantle of rigor" (Constantinou et al., 2017, p. 2) to provide the appearance of rigor that is largely unsubstantiated by researchers and left unchallenged by reviewers of qualitative studies. To some extent, this lack of transparency may reflect the absence of guidance on assessing saturation. Our review has synthesized a range of strategies that can be used by qualitative researchers to become more transparent in reporting *how* saturation was assessed, *whether* it was reached, or the *extent* to which it was achieved in a study. Researchers can now specify a strategy for assessing saturation and the criteria on which it was determined (e.g., a stopping criterion, cumulative frequency graphs, percentage of codes, code meaning). Such greater transparency has clear benefits for the rigor of individual studies but also for the quality of qualitative research as a whole. Greater transparency regarding saturation improves reproducibility of the research and raises expectations on how to report saturation, all of which move away from using generic and unsupported statements such as 'data were collected until saturation'. In addition, journals publishing qualitative research play a critical role in encouraging transparent reporting of saturation. Vasileiou et al. (2018) found that reporting of sample size justifications aligned with particular academic journals, suggesting that journal requirements may encourage norms of greater transparency in reporting saturation. Journal reviewers may also encourage transparency by asking researchers, for example: '*how did you assess saturation ?*', '*how do you know you reached saturation ? ' ,* or '*to what extent was saturation reached – in core codes, categories, meaning etc*.?'. Such requests signal that more transparent, nuanced, and rigorous reporting of saturation is expected. This should go beyond simple check-list requirements, which may simply perpetuate vague reporting that saturation was reached without study-specific details on how it was determined.

Our study has some potential limitations. We included only studies that were published in English and were outside the epistemological approach of grounded theory, and we used limited search terms for specific qualitative methods but included common methods. While these criteria may have excluded other published tests of saturation, we believe our search criteria were broad enough to capture a significant body of research on the topic. Articles identified in our review focus overwhelmingly on health research and have similar conceptualizations of saturation. While this makes the studies more comparable, these results may not be applicable to other disciplines that may conceptualize saturation differently.

5. Conclusion

Saturation is considered the cornerstone of rigor in determining sample sizes in qualitative research, yet there is little guidance on its operationalization outside of grounded theory. In this systematic review, we identified studies that empirically assessed saturation in qualitative research, documented approaches to assess saturation, and identified sample sizes for saturation. We describe an array of approaches to assess saturation that demonstrate saturation can be achieved in a narrow range of interviews (9–17) or focus group discussions (4–8), particularly in studies with relatively homogenous study populations and narrowly defined objectives. Although our systematic review identified sample sizes for saturation, we found little empirically based research to determine how specific parameters influence saturation. Further research is needed on how specific parameters influence saturation, such as the study goal, nature of the study population, sampling strategy used (i.e. inductive vs fixed sampling), type of data, saturation goal, and other influences.

Credit author statement

Monique Hennink: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization **Bonnie Kaiser:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Review and Editing, Visualization.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2021.114523.

References

Ando, H., Cousins, R., Young, C., 2014. Achieving saturation in thematic analysis: development and refinement of a codebook. Compr. Psychol. 3, 4.
Baker, S., Edwards, R., 2012. How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research. National Centre for Research Methods, Economic and Social Council (ESRC), United Kingdom.
Bryant, A., Charmaz, K. (Eds.), 2007. The SAGE Handbook of Grounded Theory. Sage, London.
Bryman, A., 2012. Social Research Methods, fourth ed. Oxford University Press, Oxford, UK.
Carlsen, B., Glenton, C., 2011. What about N? A methodological study of sample-size reporting in focus group studies. BMC Med. Res. Methodol. 11. Article 26.
Coenen, M., Coenen, T., Stamm, A., Stucki, G., Cieza, A., 2012. Individual interviews and focus groups with patients with rheumatoid arthritis: a comparison of two qualitative methods. Qual. Life Res. 21, 359–370. https://doi.org/10.1007/s11136-011-9943-2.
Constantinou, C., Georgiou, M., Perdikogianni, M., 2017. A comparative method for themes saturation (CoMeTS) in qualitative interviews. Qual. Res. 1–18.
Fofana, F., Bazeley, P., Regnault, A., 2020. Applying a mixed methods design to test saturation for qualitative data in health outcomes research. PLoS One 15 (6), e0234898.
Francis, J., Johnson, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M., Grimshaw, J., 2010. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol. Health 25, 1229–1245.
Fugard, A.J., Potts, H.W., 2015. Supporting thinking on sample sizes for thematic analyses: a quantitative tool. Int. J. Soc. Res. Methodol. 18, 669–684.

Galvin, R., 2015. How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge? J of Building Engineering 1, 2–12.
Glaser, B., Strauss, A., 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine, Chicago.
Guest, G., Bunce, A., Johnson, L., 2006. How many interviews are enough? An experiment with data saturation and variability. Field Methods 18, 59–82.
Guest, G., Namey, E., Chen, M., 2020. A simple method to assess and report thematic saturation in qualitative research. PLoS One 15 (5), e0232076.
Guest, G., Namey, E., McKenna, K., 2016. How many focus groups are enough? Building an Evidence Base for Non-Probability sample sizes. Field Methods 29 (1), 3–22.
Hagaman, A.K., Wutich, A., 2017. 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, 23–41.
Hancock, M., Amankwaa, L., Revell, M., Mueller, D., 2016. Focus group data saturation: a new approach to data analysis. Qual. Rep. 21, 11.
Hennink, M., Kaiser, B., Marconi, V., 2017. Code saturation versus meaning saturation: how many interviews are enough? Qual. Health Res. 27 (4).
Hennink, M., Kaiser, B., Weber, M.B.(, 2019. What influences saturation? Estimating sample sizes in focus group research. Qual. Health Res. 29 (10), 1483–1496.
Kerr, C., Nixon, A., Wild, D., 2010. Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. Expert Rev. Pharmacoecon. Outcomes Res. 10, 269–281. https://doi.org/10.1586/erp.10.30.
Lowe, A., Norris, A.C., Farris, A.J., Babbage, D.R., 2018. Quantifying thematic saturation in qualitative data analysis. Field Methods 30 (3), 191–207.
Marshall, B., Cardon, P., Poddar, A., Pontenot, R., 2013. Does sample size matter in qualitative research? A review of literature in IS research. J. Comput. Inf. Syst. 11–22.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Prisma, Group., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 6 (7), e1000097.
Morse, J., 1995. The significance of saturation [Editorial]. Qual. Health Res. 5, 147–149. https://doi.org/10.1177/104973239500500201.
Morse, J., 2015. Data were saturated . . . [Editorial]. Qual. Health Res. 25, 587–588. https://doi.org/10.1177/1049732315576699.
Morse, W.C., et al., 2014. Exploring saturation of themes and spatial locations in qualitative public participation geographic information systems research. Soc. Nat. Resour. 27 (5), 557–571.
Namey, E., Guest, G., McKenna, K., Chen, M., 2016. Evaluating bang for the buck: a cost effectiveness comparison between individual interviews and focus groups based on thematic saturation levels. Am. J. Eval. 37 (3), 425–440.
Nascimento, L.C.N., et al., 2018. Theoretical saturation in qualitative research: an experience report in interview with schoolchildren. Rev. Bras. Enferm. 71 (1), 228–233.
O'Reilly, M., Parker, N., 2013. 'Unsatisfactory saturation': a critical exploration of the notion of saturated samples sizes in qualitative research. Qual. Res. 13 (2), 190–197.
Rowlands, T., Waddell, N., McKenna, B., 2016. Are we there yet? A technique to determine theoretical saturation. J. Comput. Inf. Syst. 56 (1), 40–47.
Sandelowski, M., 1995. Sample size in qualitative research. Res. Nurs. Health 18, 179–183.
Turner-Bowker, D.M., Lamoureux, R.E., Stokes, J., Litcher-Kelly, L., Galipeau, N., Yaworsky, A., Shields, A.L., 2018. Informing a priori sample size estimation in qualitative concept elicitation interview studies for clinical outcome assessment (COA) instrument development. Value Health 21 (7), 839–842.
Van Rijnsoever, F.J., 2017. (I Can't Get No) Saturation: a simulation and guidelines for sample sizes in qualitative research. PLoS One 12, e0181689.
Vasileiou, K., Barnett, J., Thorpe, S., Young, t., 2018. Characterizing and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med. Res. Methodol. 18, 148.
Weller, S., Vickers, B., Bernard, R., Blackburn, V., Borgatti, S., Gravlee, C., Johnson, J., 2018. Open-ended interview questions and saturation. PLoS One 13 (6), e0198606.
Young, D.S., Casey, E.A., 2019. An examination of the sufficiency of small qualitative samples. Soc. Work. Res. 43 (1), 53–58.