Thematic Saturation In Interview-based Qualitative Research

Ross M. Woods, Worldwide University, AZ

October 2022

Researchers who conduct qualitative research with interviews need to know how many interviews are necessary, after which more interviews add no more useful information. Here's a typical case:

> Lauren was doing a qualitative research project. After all the planning and approvals, she started writing an interview questionnaire. Except for the first group of questions, which were mainly identification and demographics of interviewees, all the questions were open-ended. She asked the same set of questions in the same order, but the research plan allowed her to improvise follow-up questions so that she could explore interviewees' views on any relevant topics as they came up. She could easily use the same procedure without changes for more respondents.
>
> She did not know what would emerge from the data, and was open for her research to take to an unexpected direction if the data demanded it. Consequently, she did not predetermine all categories that she would use to analyze data. She started with a draft set, but expected to create more if the data required it; that is, data emerged during the interviews.
>
> She did not plan to have a specific number of interviews, because she did not know how many interviews she would need to draw and confirm conclusions.
>
> After each interview, she transcribed it and identified parts of the interview that were relevant to her research topic, classifying those parts according to a system of themes.
>
> The interviews varied greatly in how much new data they produced. Everything was new in the first interview, and the next several produced quite a lot of new information. After that, they didn't always follow a clear pattern; one or two might contain several items of new information, then several with little or nothing, then one or two with a little more. Later on, it was only rarely that an interviewee would say anything that hadn't already been said by a previous interviewee. Some themes had came up often, while others were quite rare.
>
> Eventually, a pattern emerged and thenall frequently-occurring items created a fairly obvious pattern that was strong enough to confirm conclusions, and Lauren thought it might be time to stop doing any more interviews. Nothing new came up any more. Should she stop doing interviews?

The problem is now over fifty years old, and was first posed by Glaser and Strauss in 1967 (Guest et al. 2020). Since then, the literature on the topic has been broad and often inconclusive.

1. Literature review

1.1 Guest et al.

Guest et al. (2006) conducted an experiment in Ghana and Nigeria following a procedure much like the example above of Lauren. Although the sampling method was purposive and nonprobabilistic (p. 62), the population was relatively homogeneous (p. 75). They asked all respondents identical questions in the same order, but explored any key responses (p. 63).

They found that, after twelve interviews, new themes emerged only infrequently and progressively so, and that definitions of their themes were also quite stable (p. 68). They used Cronbach's alpha to measure the reliability of theme frequency distribution as the analysis progressed and found that their data was internally consistent (p. 73).

1.2 Francis et al.

Francis et al. (2010) reviewed all papers published in a particular journal over a 16-month period. Eighteen mentioned data saturation and gave consistent criteria; data saturation meant that no new themes, findings, concepts or problems were evident in the data" (p. 2). Fifteen claimed to have achieved it, but it was not clear how they had done so (p. 2). Francis et al. proposed that saturation occurs when "three further interviews have been conducted with no new themes emerging" (p. 6.) They also suggested that "the analysis would ideally be conducted by at least two independent coders and agreement levels reported to establish that the analysis is robust and reliable" and that "findings ideally would be reported so that readers can evaluate the evidence." (p. 6.)

1.3 Carlsen and Glenton

Carlsen and Glenton examined 28 studies that claimed to have reached saturation, and found that 15 of them did not report convincingly how they had reached it. (2011, p. 5.)

1.4 Mason

Mason (2010) found that many institutions ignored saturation and created arbitrary norms. Based on a set of criteria, he selected 560 Ph.D. abstracts from a sample of 2,533. He found that the "most common sample sizes were 20 and 30 (followed by 40, 10 and 25)", and concluded that "PhD researchers (and/or their supervisors) don't really understand the concept of saturation." They do a large number of interviews to ensure that their sample sizes and their data are defensible. Alternatively, they use larger that necessary samples "just to be on the safe side".

1.5 Baker and Edwards

The same problem endured. Baker and Edwards (2012) attempted to resolve it by collecting articles from nineteen researchers, most of whom were established experts. It resulted in a series of well-informed but divergent responses, most of which suggested that the notion of saturation was either inappropriate or unhelpful. For example, they noted that project proposal procedures or ethics committees required researchers to state a number of interviews before research could commence. (2012, p. 6.)

It was particularly helpful that they noted Tracey Jensen's contribution about the "quality of the analysis and the dignity, care and time taken to analyse interviews, rather than quantity" and the need to "build a convincing analytical narrative based on 'richness, complexity and detail' rather than on statistical logic." (p. 5)

1.6 Flick

Flick also lists outside factors that might also determine the number of interviews:

1. The time given for the project
2. The researcher's amount of experience with qualitative research
3. Limits to the number of interview partners
4. Limitations on transcription and analysis
5. Accessibility of potential interviewees, and
6. "Resources." (2012, p. 27.)

1.7 Morse

Morse (2014) also reported the same problem several years later. Of those researchers who claimed to reach saturation, he noted that they seldom documented the methods that they used to do so (p. 558). For example, Koohestani states "As a qualitative research, there were no definite rules to determine the number of participants. Therefore, sampling was continued until data saturation so that the authors were sure no new information could be found." (2018, pp. 148f.) In this case, it seems that the researchers simply ceased interviews after an unspecified number of extra interviews that produced no new results. It was not the lack of method that might cause concern, but the lack of specific reporting and statistical justification.

1.8 Fusch and Ness

Based on earlier sources, Fusch and Ness defined saturation as being reached when "there is enough information to replicate the study ..., when the ability to obtain additional new information has been attained ..., and when further coding is no longer feasible ...." (2015, p. 1408)

Of these, replicability of the study is essential to the interview procedures, that is, the same procedure can be used for more interviewees. The reference to Guest et. al. seems to be a copyist error; they meant that saturation is reached when no additional new information can be attained (Fusch and Ness, p. 1409.) The stage at which further coding is no longer feasible is probably intended to mean that further coding is no longer beneficial. It is usually possible to divide themes into more specific themes, but is not necessarily beneficial.

Fusch and Ness thought that the concept of saturation is difficult to define, especially considering the range of different research designs, and that "data saturation for one is not nearly enough for another," mentioning ethnography as a particularly different case (p. 1408). In fact, it is for this reason that the present discussion is limited to interview-based research.

It is helpful that Fusch and Ness differentiated between the quantity and the quality of data, describing data quality as "rich ... many layered, intricate, detailed, nuanced, and more." (P. 1409) They state that "data saturation is not about the numbers per se, but about the depth of the data."

However, they then make the curious statement that: "If one has reached the point of no new data, one has also most likely reached the point of no new themes; therefore, one has reached data saturation." This is inconsistent with the distinction between the quantity and quality of data; they seem to imply that thematic saturation is the same as having enough data of high quality (p. 1409).

They also state that "There is a direct link between data triangulation and data saturation; the one (data triangulation) ensures the other (data saturation)." (P. 1411.) While the benefits of triangulation are not in doubt, it is questionable that multiple methodologies are necessary to achieve data saturation; rather the question at hand is a means to identify saturation for a particular methodology, in this case, interviews.

1.9 Galvin

Galvin examined 54 papers reporting the use of interview methods to research energy consumption in buildings (2015, pp. 2-6), but also discussed the shape of the broader literature (2015, pp. 6-8). He cited three principles that have been used to determine saturation:

1. The 'wisdom of the elders' means to follow precedents set in similar research.
2. The 'experience of the researcher' means to draw a conclusion based on the complexity of the topic and the issues involved.
3. The 'quasi-empirical foundation' is the use of Chronbach's Alpha to measure internal consistency of the data, as proposed by Guest et. al. (2006).

As Galvin correctly notes, the first two methods lack any statistical criteria for accuracy. However, his critique of the third method is less convincing. Guest et al. should not be interpreted to mean that "30 is a universal maximum for saturation." In fact, they deliberately state that "although no new themes emerged after the 30th interview, this does

not imply that 30 is a universal maximum for saturation." (p. 8) Similarly, Galvin states that the method assumes that "no more themes would be found if the total number were extended," although Guest et al. do not make that claim. Rather than erroneously claim that all themes have been found, it is better to propose that that it is relatively unlikely to find more and if one did, those themes would be outliers.

Galvin's own method is similar to that of Guest in that he also does not claim to uncover all possible themes. Galvin criticized the practice of setting an arbitrary number of extra interviews after the first interview from which no new beliefs emerged on the basis that more themes could still emerge after that. For example, Morse et. al. ceased interviews when no new themes had emerged in three consecutive focus groups (2014, p. 560). Mortazavi and Davarpanah also added "three more extra participants to produce more reliable results. (2021, p. 6.)

Perhaps more helpfully, Galvin noted that their samples were not true random samples of their target populations. Galvin also noted that meaning, interpretation, perception, and nuance are important in qualitative research, and not the way that qualitative data can be expressed in statistics (p. 6).

Probably his most useful contribution is a method for specifiying a confidence level, discussed later on.

1.10 Van Rijnsoever

Van Rijnsoever (2017) does not limit his research to questionnaires and considers different kinds of information sources and different methods, such as observations and analysis of documents and archives (p. 4). He also includes different sampling methods, such as random samples and theoretical samples. He refers to Coyne's definition of theoretical sampling, in which the researcher starts with a purposive sample and then refine methods and definitions of populations according to the most fruitful emerging avenues of inquiry and emerging theory. This includes adding more respondents in order to explore and test the theory (1997, p. 625; Miles and Huberman, pp. 29; Becker, 2017, p. 15.)[1] That is, it is an iterative approach to sampling which differs markedly from the kind mentioned above, so its pattern of identifying themes and achieving saturation could be quite different.

It is unfortunate that van Rijnsoever uses the term sub-populations to refer to triangulation. The term sub-populations is better used to refer to separate categories of respondents within a group of respondents, not a comparison of different results obtained using different methods from different kinds of data sources. (cf. p. 4). However, taken instead as a view of triangulation, his comments would probably be quite acceptable to most researchers.

1 Miles and Huberman call the evaluation of the sample "conceptually-driven sequential sampling." (1994, p. 27.)

1.11 Hennink et al.

Hennink et al. (2017) used the terms *code saturation* for thematic saturation and *meaning saturation* for "the point when we fully understand issues, and when no further dimensions, nuances, or insights of issues can be found." (p. 594). They "reached code saturation at nine interviews, but needed 16 to 24 interviews were needed to reach meaning saturation where we developed a richly textured understanding of issues. Thus, code saturation may indicate when researchers have 'heard it all,' but meaning saturation is needed to 'understand it all.'" They also found that, just because a theme occurred frequently does not mean that it was important to understanding; those that are less prevalent "may contribute equally to understanding themes in data; thus, they become important not for their frequency but for their contribution to understanding." (p. 605).

The significance of their research is that they demonstrated that thematic saturation is different from meaning saturation; it had previously sometimes been assumed that thematic saturation was sufficient to achieve meaning saturation. It also supported the conclusions of Guest et al. (2006) that a relatively small sample is adequate to reach thematic saturation.

1.12 Weller et al.

Weller et al. (2018) proposed that thematic saturation should represent only the most salient items rather than all items, and suggested that "number of unique items added by each respondent (count data) is approximately Poisson distributed." (p. 4) but found that the "negative binomial model resulted in a better fit than the Poisson ... for most full-listing examples, providing the best fit to the downward sloping curve with a long tail." (p. 6.)

They concluded that "probing and prompting during an interview seems to matter more than the number of interviews. … A small sample (n = 10) can collect some of the most salient ideas, but a small sample with extensive probing can collect most of the salient ideas." (P. 15.)

"Some domains were well bounded and were elicited with small sample sizes. Some were not. In fact, most of the distributions exhibited a very long tail–where many items were mentioned by only one or two people. ... Although the expected number of unique ideas or themes obtained for successive respondents tends to decrease as the sample size increases, this occurs rapidly in some domains and slowly or not at all in other domains." (P. 8.)

1.13 Guest et al.

With different co-authors from the 2006 article, Guest proposed a method to assess and report thematic saturation. The main idea is to compare the number of new themes emerging with the number found in the first set of interviews. They defined data saturation as thematic saturation that is, when no more themes will emerge with more interviews (2020, p. 1).

The researcher starts by determining arbitrary standards of method and rigor. First, set an arbitrary number of the first series of interviews as the "base line" (They set it at four in the example.) Second, set an arbitrary number of consecutive interviews to determine how often you will check for saturation; this is called the "run length." (He sets it at two in the

example.) Third, set an arbitrary figure as the level of lack of new information that indicates saturation. (He suggests ≤5%.)

The procedure in the second phase is as follows:

1. Find the number of unique themes in the "base line" that is, the first set of interviews.
2. Find the number of unique themes in the next series of interviews according to its run length. (In his example, it is the following two interviews.)
3. Calculate the proportion of new themes in that run length compared to the total number of themes in the base set.
4. And so on until the number of new themes in a run length meets the saturation threshold.

Its weakness appears to be that it depends greatly on some arbitrary measures that might easily have been avoided. Despite being called a "simple method," it is cumbersome in comparison to other methods.

However, it was helpful that Guest et al. noted that the curve is asymptotic (2020, p. 6), because, like other methods, it also plots a downward curve on a graph that flattens out and stops at a threshold point, which is a very low point in the curve.

1.14 General comments

The use of vocabulary in the literature is sometimes not uniform. Fusch and Ness use the term "data saturation" when they differentiate between quality of data and amount of data (2015, p. 1409), contributing to the later differentiation between meaning and thematic saturation. Van Rijnsoever uses the terms "code" and "theoretical saturation" for theme and meaning saturation respectively. (2017, p. 5.)

Similarly, "iterative" or "iteration" can refer to the addition of more informants with the consistent use of a questionnaire (Weller et.al., 2018, p. 16), but can also mean the change of method, for example, the definition of the population or the questions (Hennink et al., 2017, p. 593; van Rijnsoever, 2017, p. 4).

The method of identifying thematic saturation is only quasi-statistical. It assumes that qualitative data can be reduced to binaries (i.e., occur/does not occur), and depends on qualitative judgments to interpret qualitative data to create statistical data.

The asymptotic curve appears to address the datasets for thematic saturation discussed in Galvin's literature review (2015) and in Guest, Namey, and Chen (2020).

2. Saturation: A matter of definition

This leads to the problem of defining saturation. Is it the point at which each theme has occurred at least once, such that more interviews will probably not reveal any more? Alternatively, is saturation the point at which the researcher can already identify patterns in data and confirm conclusions, such that more interviews will probably not change those conclusions?

The idea that each theme has occurred once refers only to the range of relevant themes, not to their frequency of occurrence (Wolcott, n.d.) nor the richness of data within them. In other words, *thematic saturation* and *meaning saturation* are different concepts, and one might reach thematic saturation without reaching meaning saturation. (Hennink et al., 2017.) Theme occurrences can vary greatly in quality. Some are simply brief, mundane mentions of a theme and do not provide much useful information. In contrast, a small proportion are long, helpful explanations with great insight. Even though usefulness is a matter of individual judgment, it is still relevant to determining saturation. Put this way, the answer is that more information is necessary to achieve saturation if most data so far is not very helpful.

Another factor in saturation is the stage at which further coding is no longer beneficial. (Cf. Guest et al., 2006, p. 77; Fusch and Ness, 2015, p. 1408.) Coding need not be so fine grained that it provides no actual benefit in data analysis, although the limit of code creation is a judgement call.

The statistics of theme occurrences are not the whole story for determining saturation, because theme occurrences are simply binary. Collecting enough themes is not an assurance that one has collected enough useful data. That is, thematic saturation is inadequate by itself, because it does not indicate the quality of responses. The question is then, "Does the collected data contain enough useful information, even if more interviews probably will not uncover new themes?" Simply counting occurrences of themes reduces occurrences to a binary statistic: "From a statistical point of view the outcomes of this type of research can be classified as binary (or binomial), in that each outcome is either found or it is not found." (Galvin, 2015, p.3)

It is probably impossible to quantify the quality of qualitative data and it is at least a contradiction of epistemologies. Consequently, two things remain a matter of qualitative human judgement: determining the point of meaning saturation and determining the point at which one can confirm the patterns and conclusions that have emerged in qualitative data.

3. The nature of the curve

With more interviews, new themes become rarer until eventually no new themes appear at all. All views so far have suggested an asymptotic curve on a graph; the number of new themes follows a curve that starts high, falls steeply, and then flattens out into a long tail, reducing each time and getting closer to 0, but never reaching it.

Some researchers have sought to define the curve mathematically. The first question is whether discrete or continuous mathematics is most useful. Continuous mathematics iresults in a smooth, continuous curve that gives an exact solution that is not rounded to whole numbers. However, discrete mathematics uses whole numbers, and themes and number of interviews only exist in whole numbers.

Morse suggested an asymptotic curve (2014, p. 562) and Galvin suggests a logarithmic curve (2015, p. 11.) Weller et al. found that a negative binomial distribution was the best fit for the data they tested. (2018, p. 6.)

Another possibility is a geometric progression The binomial nature of thematic data resembles a Bernoulli trial in two ways: First, each item must be one of two options, and,

second, the order of data collection does not affect the result. The effect of the second is a common ratio for the geometric curve. Consequently, thematic data decreases at a uniform rate, which can also represent a geometric progression with a formula $y = ar^{x-1}$ where *a* is the initial value and *r* is the common ratio, which must be less that 1.0 and greater than 0.

To some extent, the mathematical definition of the curve is unimportant for two reasons. First, it is only a notional probability. Second the mathematical definition of the curve matters little as long as it is clearly asymptotic. In continuous mathematics, the number descends to a figure below 0.5, which is rounded to zero. The curve could continue until all members of the population have been interviewed, while there is still a chance that a new theme could arise. In practice, however, the number of new themes normally hits zero (no new themes arising), which is the basis of some kinds of confidence measures discussed below.

4. Why data does not follow the curve exactly

For various reasons, the curve only is a notional probability rather than an actual set of results. First, the number of interviews and the raw data of thematic occurrences can only be whole numbers. Second, a random fluctuation of only one or two creates a major aberration when number are very low. For example, the number of interviews might be as few as six and is seldom more than sixteen. The numbers at the right side of the curve are very small.

Third, real data seldom follows the curve exactly; the line usually jags erratically upward and downward to some extent. However, moving average smoothing would better represent the path of the curve; that is, one uses the average of recent interviews, say, the last three or four, instead of the raw score to plot the curve so that the curve is less jagged. Its advantages are that the pattern is clearer, outliers have less effect on trends, and in normal circumstances, the curve can be expected to more closely follow the curve. While it might indicate progress toward saturation, it would be higher risk to suggest that it can be used to predict saturation. The disadvantage of moving average smoothing is that some series of interviews reach thematic saturation at relatively low numbers of interviews, at which it is not very useful. For example, some series of interviews might reach saturation after only six interviews. (Guest et al., 2006, p. 73.)

5. Fog

Various factors are fog that inhibit the understanding and development of a thematic saturation theory.

5.1 Order of data collection

Two errors in particular indicate scenarios in which the order of data collection affects the result, that is, they do not meet the conditions of a Bernoulli trial.

Qualitative research has an “iterative” quality in which researchers refine methods and redefine definitions of themes as they go. (Hennink et al., 2017, p. 598.)

> As Matt added more interviews, he also noticed ways to improve the questions; mainly as follow-up questions to check on frequently-occurring themes and

mentions of topics that were too brief to be useful. Perhaps this helped him uncover new themes that he had missed in earlier interviews.

As Matt added more interviews, he added more themes. With the benefit of hindsight, he found that his first set of themes wasn't always as good as it had first appeared. Would it be a good idea to improve his definitions of them, and perhaps divide several of them into more specific better-defined themes? These new themes might come up less often, but he'd have more finely grained data that would result in better conclusions. It would improve his analysis, but might create confusion of the way he counted occurrences of themes. He then asked, "If I change the themes, should I go back and change all her color-coding on the interviews that I have already done? How would it affect her results?" He also wondered if he could also group several very similar themes that seldom occurred.

The most obvious kind of iteration is the way in which the researcher improves the definitions of themes during the research. Another possible cause is that the researcher might find that the nature of the problem is different from the problem formulated in the proposal and literature review.

The general solution is two pronged. First, the questionnaire needs to tested before extensive use, so that the same questionnaire will be used each time. Second, as better definitions of themes emerge, the researcher should check previous analysis to ensure that analysis of all data is consistent. The redefinition of the nature of the problem is a special case; it implies that a new research cycle is necessary with a different questionnaire.

Put another way, if researchers redefines themes as they go without rechecking past data, iterations can lead to aberrations in the curve. On the other hand, it is probable that the asymptotic curve applies when the themes are interpreted consistently. This is expressed as an hypothesis further below.

In another kind of error, sub-populations can cause aberrations in the curve. Consider this fictitious illustration:

> James ran a series of interviews in different suburbs with different demographics, although all interviewees met the criteria for the research population. The first group was in Alphaville, the second in Betaville and the third in Gammaville where socioeconomic levels sere quite different. James reached saturation in Alphaville, but then Betaville, and Gammaville each produced new different data with their own saturation points.

James's research was still quite valid as far as this interviews are concerned, but the curve was a wavy downward line and not useful for forecasting the point where it would be unlikely for more new themes to emerge. However, if he had defined each group as a different population, each could have had its own curve.

However, sub-populations are not the same as variations. Populations can differ in the amount of variation:

> Rebecca ensured that all members of the research population met the criteria for inclusion. As she did her interviews, she noticed that the data did as expected; the amount of new information gradually reduced, roughly following the curve. But then it would spike upward, and she didn't know why. Her supervisor suggested that she might be facing different sub-populations, but she didn't know what they were; she had followed the selection criteria very consistently. Alternatively, it might be a simple aberration in the data; nobody could expect the downward curve to always be smooth. In this case, the solution was simply to hold more interviews, with the possibility that the reasons for the spikes might emerge later on in the data.

This leads to the hypothesis that greater variation within a population requires a larger number of interviews. (See below in hypotheses.)

5.2 Outliers and failure to capture all themes

Galvin expressed concern that "because all themes have been found after a particular number of interviews, no more themes would be found if the total number were extended. … A theme with such a low frequency of occurrence may seem trivial, but in issues to do with social justice ... it is essential to bring the marginal cases to light."(2105, p. 8) However, the number of interviews indicates only the probability that few new themes would emerge, and if they did, they would be outliers. The only way to ensure that all themes would emerge, as Galvin hoped, is to interview all members of the population.

When outliers are important, Consider this fictitious example:

> Brad was studying a particular disease that affected most people quite mildly, but he was troubled by outliers, that is, themes that rarely occurred. A small number of patients needed to be hospitalized, and a few even died. These cases were important to the final research even though they were were outliers and did not occur often enough to form a pattern or to even understand much about them.

In fact, Galvin's own method did not expect to reveal all relevant themes. Outliers are by definition statistically insignificant. When they are important, it is a matter of defining the specific populations. A group of important outlier cases be treated them as a different population in new research with a new research question in order to find its particular characteristics.

5.3 Could a theme be there but not found?

Another issue is whether a theme is out there but is not found. Its importance is that, if it exists but is not captured, then the results are skewed.

Van Rijnsoever expressed concern about the risk that data exists but is not captured, and even concludes that the "mean probability of observing codes is more important than the number of codes in the population for reaching .. saturation." that is, when "all the codes in the population have been observed once in the sample." (van Rijnsoever, 2017, pp. 1, 5f..) However, it is unclear about how one could calculate the probability that data was not captured; one would have to know what data was not captured in order to work out such a probability. The researcher works only with the data (that is, the transcripts of interviews), not speculation that a theme might not have been captured. Having said that, the researcher needs good quality data.

The method of thematic saturation also assumes that a theme will appear in the data if it occurs in the sample. Some subjects are taboo, such as those that subjects consider to be distasteful or embarassing, so interviewees might be reluctant to discuss them. This is not an argument either for or against treating themes as binary information. It is the same whether one was looking for countable themes or was an ethnographer dealing only with qualitative data; identifying those themes still depends on the ability of the researcher to explore and to uncover relevant information.

5.4 Random sampling

Galvin expressed a preference for a "true random sample" of the population. (2015, pp. 4,16) It is true that a fully randomized sample is mathematically ideal. However, it is often not possible when conducting interview research. Even when the population is carefully defined, it might not be possible to know the total number of members the population has from which a random sample could be made. However, it is hypothesized that the closer the selection and order of interviewees approaches a randomization, the more likely the sample is to be representative.

5.5 Purposive sampling

Purposive sampling is generally defined as the researcher's selection of subjects. The method is ambiguous in that it can range from almost fully randomized to something more like a snowball or opportunistic sample. Galvin is probably correct to suggest that "the degree of non-randomness is usually unknown." (2015, pp. 16.)

Although many writers refer to purposive sampling, it is actually an umbrella terms for up to sixteen different kinds of sampling (e.g. Guest et al., 2006, p. 61; based on Patton, 2002). That is, the term is imprecise if used without further qualification. For example, researchers can define a population one way but select respondents from a group defined differently. Some so-called "sampling techniques" are really the specification of particular kinds of populations:

1. Selecting more knowledgeable respondents is the same as defining the population as persons who are knowledgeable about the topic being researched (and who meet any other specific criteria).

2. Selecting atypical or critical cases is the same as defining the population as persons with demographic characteristics who have had atypical or critical experiences of the topic being researched (and who meet any other specific criteria).
3. Selecting politically important cases is the same as defining the population as persons whose experiences of the topic being researched are either politically significant or politically benign (and who meet any other specific criteria). (Miles and Huberman, 1994, p. 28f; Morse, 2014, p. 229.)

5.6 Researcher expertise

Mason (2010) mentioned that inexpert researchers might create sets of themes that are not helpful. For example, in a study of obesity-related stigma, all subjects might mention stigma related to their obesity and the theme would soon become saturated. A more fine-grained set of categories would result in better research that produce a better understanding of the phenomenon. (Charmaz, 2006, p. 114.) Even so, occurrences of poorly chosen themes would still follow the asymptotic curve as long as other conditions are met.

5.7 Ethnography

Some mentioned the difficulty achieving thematic saturation in ethnography. (e.g. Fusch and Ness, 2015, pp. 1408; Miller, 2012, p. 31.)

However, ethnography is not appropriate to mapping as an asymptotic curve. Ethnography is highly iterative, perhaps the extreme case. Ethnographers do not consistently use the same set of questions, and the starting questions are different from those asked later on in the study. The same informant might be interviewed several or even many times. In other words, ethnography is inconsistent with the interview methodology that presumes interviewees will be interviewed only once and that all interviewees will be asked the same questions.

Ethnographers do not presume that communities are homogeneous; part of their role is to explore the relationships between community sub-groups. Even though many ethnographers include some ancillary demographic statistics, ethnographic data is so qualitative that it is not amenable to expression in statistical form.

Moreover, the number of themes can increase throughout the study, so thematic saturation is not always a helpful guide as to when to stop. Ethnographers sometimes ask, “How can I stop when have am still facing an increasing number of new themes?” They seldom say they can stop interviewing because they have no more new themes. Ethnographers stop when they have collected data to confirm conclusions that answer their research questions.

If ethnographic interviews could be analyzed thematically, many of the more important themes do not emerge early in the study and the mapping of themes might form a very different kind of curve.

5.8 Unusual cases

Several unusual cases (Baker and Edwards, 2017a, p. 5) still follow the asymptotic curve. A single question in a single interview is enough to establish whether something is possible.

When the population comprises only one interviewee, the curve is truncated after the first interview, and the number of themes in the first interview is the total number of themes. Very small populations with only two, three, or four members are similar. The curve is truncated after the interview with all members of the population, and the number of themes in the interviews is the total number of themes. Very small populations, however, require qualification to the formulas:

$$m = n + e$$
$$m = n.p$$

For example, *e* equals 0 when the total population is 1. That is, one cannot conduct any more interviews because the only population member has already been interviewed. Similarly, *e* equals 1 when the total population is 2. If *e* is an arbitrary measure of rigor, the upper limit of *e* is the number of population members – the number of interviews. The rule for p is similar; the upper limit of n.p is number of population members – the number of interviews.

6. When to stop and confidence levels

Then the question arises, "How can I know which interview will be the last to provide new information for thematic saturation?" Guest, Namey, and Chen (2020) suggest that *thematic saturation* is reached when an interview reveals no new themes at all. However, this carries some risk, because one interview might by coincidence be an outlier. In other words, that specific method contains no measure of confidence level.

In this sense, a confidence level is a measure of assurance that the data collection is complete and indicates the minimum size of the population needed to achieve thematic saturation. Put another way, it represents the amount of confidence that one can place in the accuracy of the results, allowing for a margin of error.

Below is a series of methods of setting confidence levels. They all involve an arbitrary whole number to indicate rigor and a method of calculation, both of which should be included in any methodology statement. In some methods, a greater number indicates greater probability that no new themes will emerge with more interviews.

First, Guest, Namey, and Chen (2020) suggest that thematic saturation is reached when the number of new themes in an interview is 5% or less of the total number of themes found in all previous interviews. The figure of 5% is pragmatic but arbitrary. For example, a less rigorous study might set the figure at 10% while a more rigorous study might set the figure at 2%. This option has the advantage of using a figure to define the rigor of the study.

A second alternative is to hold an arbitrary number (*e*) of more interviews that produce no more new information. The researcher might say could say, "I've done enough if I've held three more interviews with no new results." The higher the value of *e*, the greater the

rigor.[2] This results in a simple formula: $m = n + e$ where m is the minimum number of interviews after which no new information is probable and n is the number of interviews conducted. (E.g. Morse et. al., 2014, p. 560; Mortazavi and Davarpanah, 2021, p. 6.)

Nobody is saying that this method has failed; the critique is that it is only a rule of thumb and has no theoretical basis. (Cf. Galvin, 2015, pp. 7f.) However, this second measure is not a theoretical orphan because it is consistent with an asymptotic curve. As the curve approaches zero, it eventually reaches a value less than 0.5 (allowing for rounding to the nearest whole number), and that subsequent values will be even closer to 0. In other words, it is probable and normal that more interviews will not yield new themes. If they do occur, they are outliers.

A third approach is to vary the number according to the number of interviews conducted. The number of extra interviews could be an arbitrary proportion of the number of interviews already done; the researcher could say, "If I've held ten interviews with no new results, then I'll add 20% more (another two interviews), and it will be enough unless I find something new." This results in the formula: m = n.p, where p is the arbitrary proportion, and n is rounded up to the nearest whole number.

Galvin proposes a fourth approach (p. 11). which is a method for calculating the number of interviews (n) necessary to have a stated level of confidence (P) that all the relevant themes that are held by proportion (R) of the population will occur within the interview sample. It assumes that the interviewer will be able to identify themes when they occur. For example, a researcher needs 29 interviews to be at least 95% confident that the sample has included all the issues occurring in 10% or more of the population. In some ways, this method is quite similar to others in that it uses arbitrary numbers (95%, 10%) as measures of rigor. On a graph, it follows a logarithmic scale (p. 11):

$$n = \frac{\ln (1 - P)}{\ln (1 - R)}$$

As a confidence level, it indicates that a theme will emerge within the interviews conducted, so that the researcher can cease interviewing as soon as the requisite value for n is reached.

Implications

This leaves the method of achieving thematic saturation to be highly qualified:

1. It only indicates the number of themes, not the richness of data.
2. It depends on well-defined themes.
3. It depends on the extent of probing in interviews.
4. It depends on using the same questions each time in the same order.
5. It depends on interpreting all answers the same way.

2 When e equals 1, this is the same as Guest, Namey, and Chen's option of "no new themes." However, the rigor is greater when e is greater than 1.

6. It depends on a fairly homogeneous population, even if there is considerable variation with in it (as opposed to identifiable sub-populations).
7. It depends on arbitrary definers of rigor.
8. It does not guarantee that all themes will be identified in the sample, although any themes not identified are almost certainly outliers.
9. It does not guarantee that all data collected is rich enough to achieve meaning saturation.

Researchers have varied greatly in describing the interview method to which thematic saturation applies. The first implication of the above discussion is the need to collate and clarify aspects of the method as developed so far:

1. Write the interview questions and test them before extensive use so that you can ask all interviewees the same questions in the same order (Fusch and Ness, pp. 1409f., Guest et al., 2006, p. 63).
2. In the interviews, ask probing and prompting questions that create full opportunity to capture data. (Weller et al, 2018, p. 15; Guest et al., 2006, p. 63.)
3. Create a fine-grained set of themes that will result in a better understanding of the phenomenon. (Mason, 2010; Charmaz, 2006, p. 114; Guest et al., 2006, p. 77.)
4. If better definitions of themes emerge, check the previous analysis to ensure that analysis of all data is consistent.
5. Use two independent coders to analyse data and report their agreement levels (Francis et al., 2010 p. 6) This isuggests best practice although it is not feasible for lone researchers.
6. Report the details of methods and findings so that readers can evaluate the evidence. (Francis et al., 2010, p. 6.)
7. If the nature the research problem needs to be redefined during research, it might result in a new research cycle with a different questionnaire and a different progression toward thematic saturation.
8. Saturation occurs when:
    a. the researcher has done a series of consecutive interviews (according to the confidence level, e.g. three interviews) with with no new themes or qualitative information emerging." (Francis et al., 2010, p. 6.)
    b. further coding is no longer beneficial. (Based on Fusch and Ness, 2015, p. 1408.)

Thematic saturation has several other interesting implications. First, thematic saturation is not necessarily a guide as to when to stop conducting interviews, because the data collected so far might not be rich enough to achieve meaning saturation.

Second, the principle of the asymptotic progression applies whether or not the researcher specifically counts themes. Consequently, if a researcher collected data that was rich enough to support conclusions, he/she might not need to count themes and monitor progress toward thematic saturation. While careful tracking of themes is still helpful to systematize analytical procedures in large research projects, meaning saturation is the main event and thematic saturation is a sideshow.

Outliers take two forms. Some are so infrequent that they do not occur in the sample, and are mathematically defined by the measure of rigor used to define the sample size. The other kind of outlier is those that are captured in the sample, but do not occur frequently

enough to be significant. For these to be defined mathematically a different measure of rigor needs to be defined.

The method above suggests that the sample of interviewees adequately represents of the population, although only in terms of the number of themes relevant to the particular research question and subject to the arbitrary definers that specify the confidence level in the data. This suggests a formula for calculating statistically valid representative samples. If no more interviews would produce new themes, then the data already gathered represents the whole population. Next, one does not need to know how many members in the population to be able to take a representative sample. Last, sample sizes are much smaller with this method than those required for random samples using current formulae.

7. For further research

Researchers so far have used three kinds of methods to support their views: mathematical theory (Galvin, 2015), simulation (van Rijnsoever, 2017), and real surveys in the field (Guest et al., 2006). Would it be helpful and possible to modify data theory so that these three kinds of data are consistent?

If there were one data theory (or even competing theories), it would be possible to express it as an algorithm so that at least some aspects of data analysis could be computerized.

So far, these methods of identifying the size of the sample for identifying themes are limited to one particular application, that is, a series of interviews. However, this is a particular kind of dataset, so it follows that it has properties from which further applications could be made. What is the definition and statistical properties of this dataset, and what broader implications does it have? What is the implied formula for calculating statistically valid representative samples?

Would there be benefit is applying thematic saturation to other kinds of data collection methods? Guest has already mentioned its use for focus groups and observations, probably with no change. (Guest, 2020, p. 6; Morse et. al., 2014, p. 560.) What would be the shape of the curve for other methods of data collection, such as theoretical sampling (Van Rijnsoever, 2017), ethnography, and literature analysis? (Guest, 2020, p. 6, Fusch and Ness, 2015, pp. 1408.) Could the method be appropriately applied in quantitative research?

What is the relationship between Galvin's formula above and the method of "three consecutive interviews"? Are they consistent or even equivalent versions of the same underlying concept?

The past literature has also generated several hypotheses that still need testing, many of which can be expressed as statements of direct proportion if the data is quantitative:

1. Thematic saturation always precedes meaning saturation.
2. Van Rijnsoever's (2017) hypotheses on theoretical sampling seem to apply equally well to purposive sampling with a Bernoulli trial. He suggests that "the larger the number of codes [i.e. themes] ..., the more sampling steps are required to observe them all." (p. 5.) This correlation is not at all certain; rather the wider variation of views, the greater the sample must be to observe them.

3. Van Rijnsoever's (2017) second hypothesis, which is very plausible, is that "the more often a code is present in the population, the larger are the chances that it will become observed."
4. Van Rijnsoever's (2017) third hypothesis is that the more likely a theme occurs and is observed, the smaller the sample needed to reach saturation. This is very similar to Guest et al. (2006).
5. Even when all respondents meet the criteria of the target population, it is unlikely that target populations are completely homogeneous and equally unlikely that respondents are always consistent in understanding interview questions. Put another way, a sample is not necessarily homogeneous just because the researcher strictly enforces the selection criteria for respondents:
    a. The greater the homogeneity, the more it is likely that all responses will be similar. In the same way, the more variety in the target population, the more likely that responses will be different.[3]
    b. The less homogeneous a population is (i.e. the greater the variation), the more interviews it will probably take to reach both thematic and meaning saturation; that is, the flatter the curve. (Guest et al., 2006, p.79; (Hennink et al., 2017, p. 606.)
    c. When interviewees tend to interpret themes in the same way, their responses will tend to follow an asymptotic curve.

8. Bibliography

Baker and Edwards. (2012). "Introduction" in Baker, Sarah Elsie and Rosalind Edwards (eds.) "How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research." *National Centre for Research Methods Review Paper*, pp. 3—6.

Carlsen, Benedicte and Claire Glenton. (2011). "What about N? A methodological study of sample-size reporting in focus group studies." *Medical Research Methodology*, 11:2.

Charmaz, Kathy (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA: Sage. Cited in Mason, Mark. (2010). "Sample Size and Saturation in PhD Studies Using Qualitative Interviews" *Forum: Qualitative Social Research* Volume 11, No. 3, Art. 8. September.

Coyne, Imelda T. (1997). "Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries?" *Journal of Advanced Nursing*, 26, 623–630.

Denzin, Norman K. and Yvonna S. Lincoln (eds.) (1994). *Handbook of Qualitative Research.* Thousand Oaks, Ca.: SAGE Publications.

Flick, Uwe. Baker, Sarah Elsie and Rosalind Edwards. (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 Review Paper.*. P. 27.

3 These refer only to likelihood; it is possible that populations are heterogeneous but responses are similar, and populations might be homogeneous but responses are not.

Francis J.J., Johnston M., Robertson C., Glidewell L., Entwistle V., Eccles M.P., Grimshaw J.M. (2010). "What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health*, 25:10, 1229-1245.

Fusch, P. I., and Ness, L. R. (2015). "Are We The re Yet? Data Saturation in Qualitative Research." *The Qualitative Report*, 20(9). 1408-1416. Retrieved from https://nsuworks.nova.edu/tqr/vol20/iss9/3

Galvin, Ray. (2015). "How many interviews are enough? Do qualitative interviews in building energy consumption research produce reliable knowledge?" *The Journal of Building Engineering*. (Just Solutions Cambridge Working Paper 2014B).

Guest, Greg, Arwen Bunce, and Laura Johnson. (2006). "How many interviews are enough? An experiment with data saturation and variability." *Field Methods*, 18 (2006) 59-82.

Guest, Greg, Emily Namey, and Mario Chen. (2020). "A simple method to assess and report thematic saturation in qualitative research." *PLoS ONE* 15(5):e0232076. https://doi.org/10.1371/ journal. Pone.0232076.

Hennink, Monique M., Bonnie N. Kaiser , and Vincent C. Marconi. (2017). "Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough?" *Qualitative Health Research*. Vol. 27(4) 591 –608.

Howard S. Becker. (2017). in Baker, Sarah Elsie and Rosalind Edwards. (201a). "How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research." *National Centre for Research Methods Review Paper.* P. 15.

Koohestani, H. R., Soltani Arabshahi, S. K., & Ahmadi, F. (2018). "The paradox of acceptance and rejection: the perception of healthcare professional students about mobile learning acceptance in Iran University of Medical Sciences." *Qualitative Research in Education*, 7(2), 144-169. doi:10.17583/qre.2018.3341.

Mason, Mark. (2010). "Sample Size and Saturation in PhD Studies Using Qualitative Interviews" *Forum: Qualitative Social Research* Volume 11, No. 3, Art. 8. September.

Miller, Daniel in Baker, Sarah Elsie and Rosalind Edwards. (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 Review Paper.*, p. 31.

Mores, Janice M. (1994). "Designing Funded Qualitative Research" in Miles, Matthew B. and A. Michael Huberman. (1994). *An Expanded Sourcebook: Qualitative Data Analysis. Second ed. Thousand Oaks, Ca. Sage Publications.* Pp. 220-235.

Morse, Wayde C., Damon R. Lowery, and Todd Steury. (2014). "Exploring Saturation of Themes and Spatial Locations in Qualitative Public Participation Geographic

Information Systems Research, *Society & Natural Resources: An International Journal*, 27:5, 557-571, DOI: 10.1080/08941920.2014.888791.

Mortazavi, M. and Davarpanah, A. (2021). "Implementation of a Thematic Analysis Method to Develop a Qualitative Model on the Authentic Foreign Language Learning Perspective: A Case Study in the University of Northern Cyprus." *Education Sciences*, 11, 544. https://doi.org/10.3390/educsci11090544.

Patton, M. (2002). *Qualitative Research and Evaluation Methods*. 3rd ed. Thousand Oaks, CA: Sage. p. 61. Cited in Guest, Greg, Arwen Bunce, and Laura Johnson. (2006). "How many interviews are enough? An experiment with data saturation and variability." *Field Methods*, 18 (2006) 59-82.

van Rijnsoever, Frank J. (2017). "(I Can't Get No) Saturation: A simulation and guidelines for sample sizes in qualitative research." *PLoS ONE* 12(7): e0181689. https://doi.org/10.1371/journal.pone.0181689.

Weller S.C., Vickers B., Bernard H.R., Blackburn A.M., Borgatti S., Gravlee C.C., Johnson, J.C. (2018). "Open-ended interview questions and saturation." *PLoS ONE* 13(6): e0198606. https://doi.org/10.1371/journal.pone.0198606.

Wolcott, Harry. (N.d.). Cited in p. 4. Baker and Edwards. (2012). "Introduction" in Baker, Sarah Elsie and Rosalind Edwards (eds.) "How many qualitative interviews is enough? Expert voices and early career reflections on sampling and cases in qualitative research." *National Centre for Research Methods Review Paper*, pp. 3—6.