Practice Article

EJA | EVALUATION Journal of Australasia

How many interviews or focus groups are enough?

Evaluation Journal of Australasia
2024, Vol. 24(3) 211–223

DOI: 10.1177/1035719X241266964
journals.sagepub.com/home/evj
Sage

Kizzy Gandy
Verian, Australia

Abstract

When it comes to qualitative evaluation data, is more always better and what determines value for money? This article proposes two steps for evaluators and those responsible for procuring evaluations to answer the question 'how many interviews or focus groups are enough?' Step 1 is to consider the nature of the evaluation question to determine the sampling goal, importance of thematic saturation, and an appropriate sampling strategy. The article provides guidance on how many interviews and focus groups are needed to achieve different levels of thematic saturation based on empirical tests in the published literature. Step 2 is to check the skills of the evaluator, including whether they integrate behavioural science into their discussion guide and analysis to mitigate bias. This will determine – regardless of the number of interviews and focus groups – whether they will be able to generate useful insights for decision-making from the data. The article concludes that it is not sufficient to assess an evaluation plan's value for money by sample size alone and consideration also must be given to the characteristics of the evaluation design and the skills of the evaluators undertaking the project.

Keywords

qualitative sample size, focus groups, interviews, value for money, thematic saturation

Corresponding author:
Kizzy Gandy, National Director, Program Evaluation, Verian, 320 Pitt Street, Sydney, NSW 2000, Australia.
Email: kizzy.gandy@veriangroup.com

What we know already

- Principles, guidelines, and tools for choosing a suitable sample size in qualitative research are debated, and justification for sample size sufficiency is poorly reported in published research across a range of disciplines.
- Lack of conceptual clarity about defining sample size sufficiency results in critiques of qualitative samples being 'too small' without adequate justification.
- Defence of sample size is most frequently supported with reference to the principle of thematic saturation and to pragmatic considerations (e.g. time constraints). However, the goal of 'thematic saturation' is overly simplistic if it is not situated within a more encompassing assessment of data adequacy; and replacing an assessment of sample size sufficiency with 'pragmatic considerations' only serves to perpetuate sample size norms and rules of thumb that lack validation.

The original contribution this article make to theory and/or practice

- The question of what sample size is needed for qualitative research is frequently asked by evaluators and those who commission evaluations but not frequently discussed in the literature. By encouraging transparency and understanding of how sample size and other factors affect the reliability and validity of evaluation findings, this article support evaluators and those commissioning evaluations to develop a shared understanding of the markers of quality in qualitative evaluation data.
- The existing published evidence on qualitative sample size sufficiency primarily focuses on research rather than evaluation. This article answers the question of 'How many interviews or focus groups are enough?' in relation to the unique conditions of evaluation.
- It is implicitly assumed in thematic saturation studies that saturation is always important. This is true for generating generalisable knowledge, but evaluators tend to prioritise program-specific performance insights. Generalisable knowledge is associated with a positivist paradigm whereas evaluators typically move between positivist, postmodern, and constructivist paradigms across different key evaluation questions and draw on multiple sources of data. Therefore, qualitative data is typically used by evaluators to develop a depth of understanding rather than breadth, and sometimes qualitative sample sizes as low as one can be justified.
- The article presents a new categorisation system to identify when thematic saturation is more or less important in evaluation based on the author's own experience of overseeing more than 70 evaluations over many years.

Introduction

It is rare in evaluation that we can ever study the whole population of interest, so we usually have to limit our data collection to a sample of the population. This article aims to support both evaluation consultants who must propose qualitative sample sizes in their evaluation plans, and those responsible for procuring evaluations who must judge whether an evaluation plan represents value for money. It answers the question 'how many interviews or focus groups are enough?'

To define an adequate *quantitative* sample size before heading into the field, we can often rely on statistical calculations such as power analysis for hypothesis testing (the association between two variables) or margin of error for sample representativeness (the prevalence of an outcome). Most of the time, the larger the sample the better (Martínez-Mesa et al., 2014).

For *qualitative* data, defining an adequate sample size is less straight forward. Many qualitative evaluators rely on rules of thumb – anywhere from 5 to 50 participants (Dworkin, 2012) – while others aim to maximise the sample size within the project's available time and budget.

But is more always better and what determines value for money when it comes to qualitative evaluation samples?

Data collection represents value for money when it provides reliable and valid evaluation findings

Reliable means the findings are stable or consistent across samples. For qualitative data, 'thematic saturation' – where all key themes have been discovered – is often considered critical for reliability. However, reliability is also affected by the skills of interviewers to elicit responses in a consistent manner.

Valid means the instrument used to collect the data (e.g. discussion guide) is measuring what it says it's measuring – so we can interpret the findings accurately. The extent to which evaluators design data collection instruments based on theoretically sound concepts, and can elicit unbiased responses, is critical to generating valid data.

It is helpful to think through two key steps to determine whether qualitative data represents value for money

The first step is to consider the nature of the evaluation question. This will determine whether you even need to collect new, self-reported, qualitative data, and whether that data should come from interviews or focus groups. It will also determine the sampling goal, importance of thematic saturation, and an appropriate sampling strategy. If thematic saturation is important, this article summarises the empirical research that can help you determine how many interviews or focus groups are enough.

The second step, which is often overlooked, is to check the skills of the evaluator. This will determine – regardless of the number of interviews and focus groups – whether they will be able to generate useful insights for decision-making from the data.

Step 1: What is the nature of the evaluation question?

There are many types of evaluation questions which can broadly be categorised as:

1. Appropriateness evaluation questions, for example, To what extent does the program address an identified need?
2. Process evaluation questions, for example, Was the program implemented as intended?
3. Outcome evaluation questions, for example, Did the program achieve its intended outcomes?
4. Economic evaluation questions, for example, Is the program cost-effective? (Better Evaluation, 2022).

The nature of the evaluation question should determine the most suitable evaluation design (experimental, observational, theory-based, and economic), which in turn should determine the source and type of data needed. So, before you start thinking about sampling, it's important to briefly go back a step, and check that you actually have a need for new (vs. existing) data, self-reported (vs. administrative) data, and qualitative (vs. quantitative) data. The evaluation question also informs the choice of interviews versus focus groups. See Box 1.

Box 1: Evaluation planning

Evaluation data should be aligned to the evaluation question under investigation. In the planning stage of an evaluation, key considerations are:

Evaluation design

Experimental designs use a counterfactual to make rigorous claims about causality and are therefore suited to answering outcome evaluation questions.

Observational designs answer questions based on what the evaluator observes, including associations between a program and outcomes, making them suited to answering appropriateness and process evaluation questions.

Theory-based designs can demonstrate a program contributed to outcomes by providing explanatory evidence for observed change, making them suited to answering appropriateness, process, and outcome evaluation questions.

Economic designs use economic analyses such as Cost–Benefit Analysis to answer economic evaluation questions and typically rely on estimates of effect size from an experimental design.

Source of data

New versus **existing data:** For ethical and efficiency reasons, we should try to minimise the data collection burden on respondents. Therefore, if existing data are available to answer our evaluation question, it should be prioritised over new data collection.

Self-reported versus **administrative data:** For both quantitative and qualitative data, it is important to consider whether self-reported data (e.g. asking people how much they earn), as opposed to administrative data (e.g. tax return data), will produce valid findings. Self-reported data risks response bias when it comes to measuring behaviour, but administrative data may not be available to answer questions about how people think or feel.

Type of data

Qualitative versus **quantitative data:** If we need to collect new data, we should consider what we can learn from qualitative versus quantitative data. Quantitative data are best suited to estimating the effect size of a program. Qualitative data are best suited to understanding participants' experience of a program, including barriers and enablers to achieving outcomes.

Focus group versus **interview:** Another question to be asked during the planning stage is: what is the advantage of a focus group versus an interview? Focus groups are not just a method to add numerical weight to the project. They stimulate discussion between participants to surface ideas that may not have occurred outside the group. When the group is homogenous, focus groups can reduce inhibitions. They are best used to explore issues where there is community debate or to identify social norms (Cleary et al., 2014). A randomised study of qualitative data collection methods found that sensitive themes are volunteered more frequently in focus groups compared to interviews, but interviews are more efficient than focus groups because they generate more themes per participant (Guest, Namey, Taylor, et al., 2017).

Type of experience question

If you've confirmed that to answer your evaluation question, you need to collect new qualitative data, you probably have an appropriateness or process evaluation question. These questions are often concerned with the experiences of individuals who interact

with a program. By identifying whose experience matters, and in what ways, we can logically identify an appropriate sampling goal.

From the author's own experience of undertaking evaluations over many years, there are three main types of experience questions: (1) what is the average experience, (2) how can we mange risks to positive experience, and (3) are there any unusual experiences we can learn from?

Sampling goal and sampling strategy

Your **sampling goal** will be to achieve certain sample characteristics that help answer your experience question. If your question is about the average experience, your goal may be to achieve a sample that is representative of the wider population. However, if your question is about unusual experiences, your goal may be to achieve a sample of information-rich cases.

Your **sampling strategy** is how you will select your sample to achieve your sampling goal.

Probability sampling involves random selection so every person in the population has equal chance of being selected.

Non-probability sampling includes:

- Purposive (information-rich participants are selected based on theory or prior insight).
- Snowball (initial participants are selected and asked to help recruit other participants).
- Quota (participants with specific characteristics are selected until a pre-determined number is reached).
- Convenience (participants are selected due to convenience).

Thematic saturation

The criterion most frequently used to judge qualitative sample size sufficiency is thematic saturation (Vasileiou et al., 2018). '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' (Hennink & Kaiser, 2022, p. 2).

Not all qualitative data collection needs to achieve saturation to the same extent to generate reliable and valid findings (Boddy, 2016). The relative importance of achieving thematic saturation from qualitative data collection depends on the sampling goal. For example, if you seek to identify learnings about the program design from unusual experiences, saturation is less important than if you seek to understand average experiences. Therefore, we should prioritise the question 'how much saturation is enough?' over 'has saturation occurred?'

When saturation is important, probability sampling is a very inefficient approach if the probability of observing themes in the population (e.g. unusual experiences) is low (van Rijnsoever, 2017). Purposive sampling is often the most efficient for reaching saturation across all qualitative sampling strategies because it allows information-rich cases to be selected based on theory or prior insight.

Table 1 shows that the type of experience question you seek to answer (step 1) will determine the sampling goal (step 2). The importance of thematic saturation to the sampling goal (step 3) will then determine the sampling strategy (step 4). Box 2 provides a reminder that strategy without flexibility may reduce value for money.

Box 2. Being pragmatic and adaptive to achieve value for money

Ideally, as evaluators we will always have a rigorous sampling strategy, but at times, we have to be pragmatic and take whatever sample we can get, given the limitations of the real world.

Equally, often in evaluation procurement, a specified number of interviews are required. This works well for comparability of pricing, but there ought to be flexibility once a contract is signed for the evaluator and funder to work together – applying all the principles in this article – to decide on what is really the right scale and scope to achieve value for money.

Sample size

After deciding on a sampling strategy, you are ready to consider how many interviews or focus groups you need. If thematic saturation is important, there is empirical research and mathematical/statistical models to help you. Most of the evidence relates to emergent thematic analysis (where the evaluator uncovers new themes in the data to build theory, such as barriers and enablers to program participation that weren't anticipated) so we have to assume that the findings apply equally to framework analysis (where the evaluator looks for prespecified themes in the data to test a hypothesis or validate the program's Theory of Change).

A recent systematic review of 23 articles that conducted empirical tests of saturation in qualitative research found that on average the datasets reached 90% saturation between 12–13 interviews and 5–6 focus groups. The lowest sample sizes for 90% saturation were 5 interviews or 1 focus group, and this was associated with relatively homogenous study populations and narrowly defined objectives (Hennink & Kaiser, 2022).

Table 1. Link between type of experience question, sampling goal, importance of saturation, and sampling strategy.

Step 1	Step 2		Step 3	Step 4	
Type of experience question	Sampling goal	Examples	Importance of thematic saturation to sampling goal	Sampling strategy	Examples
What is the average experience?	Representative	All population sub-groups are included	Very important	Probability sampling or non-probability sampling	Purposive/ Quota
How can we mange risks to positive experience?	Extreme or edge cases	Focus on vulnerable groups (assumes you have knowledge about the population)	Moderately important	Non-probability sampling	Purposive/ Snowball
Are there any unusual experiences we can learn from?	Unexpected cases	Identify positive deviance (assumes you have knowledge about the population)	Not important	Non-probability sampling	Purposive/ Convenience

But what if 90% saturation is not necessary to answer the evaluation question? For evaluation questions about risks or opportunities to improve stakeholder experience, somewhere in the range of 50–80% saturation may be sufficient. This will be a judgement call.

Furthermore, evidence shows that most novel information in a qualitative dataset is generated early in the process, and generally follows an asymptotic curve, with a relatively sharp decline in new information occurring after just a small number of data collection/analysis events (Guest et al., 2020). Therefore, some evaluations may require only very small qualitative samples. For example, Figure 1 shows that the number of new codes (themes) that emerged from an inductive thematic analysis of 60 in-depth interviews among female sex workers in West Africa declined significantly after the first six interviews. Likewise, Figure 2 shows that from 40 focus groups with African-American men in North Carolina on the topic of health-seeking behaviour, the majority of themes were identified within the first focus group, and more than 80% of all themes were discoverable within two to three focus groups.

One saturation study (Guest et al., 2020) applied the bootstrap method[1] to three existing qualitative interview datasets. For each dataset, the researchers generated 10,000 resamples from the original sample. They found that six was the median number of interviews needed to reach 80% saturation, and no more than 5% of new information

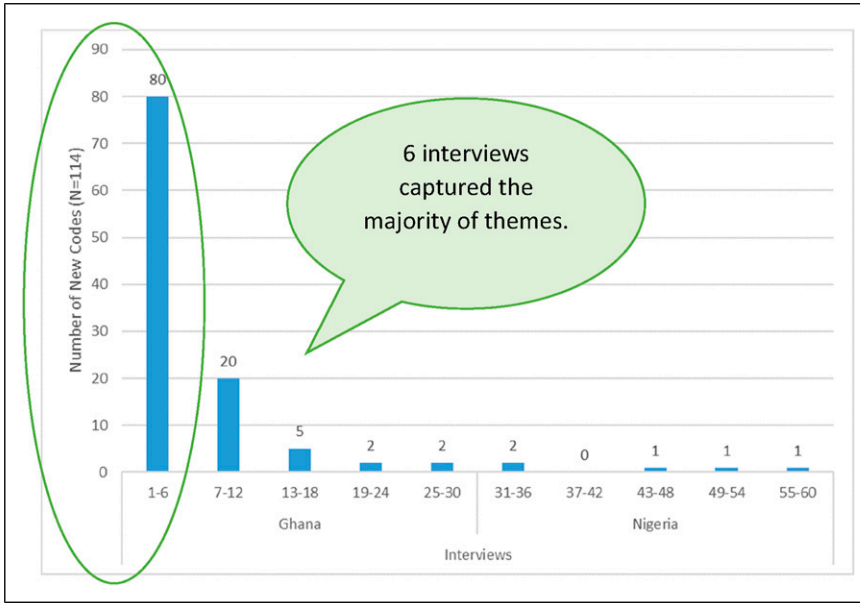

Figure 1. Achieving thematic saturation with interviews (Source: Namey, 2017).

was contributed by interviews seven and eight. So the question becomes whether these additional interviews represent value for money?

Ultimately, determining and justifying sample sizes for qualitative research is less about numbers (n's) and more about the ability of the data to provide a rich and nuanced account of the phenomenon studied. Therefore, value for money could be improved even after procurement has been completed by stopping data collection when no new themes are found. This is where it helps to refocus the routine practice of criticising qualitative research for 'small' sample sizes (Hennink & Kaiser, 2022), to instead ask about the skills of the evaluator.

Step 2: What is the skill level of the evaluator?

The purpose of evaluation is to inform decision-making. When it comes to qualitative evaluation data, insights for decision-making are largely a function of quality over quantity. And quality is largely a function of the skills of the evaluator.

If a qualitative evaluator is not trying to measure outcomes, but rather understand experiences or the reasons why outcomes did or didn't occur, then an experienced interviewer, with a clearly defined research topic, and a small number of well-selected homogeneous interviewees (with adequate exposure to the program) can produce highly relevant information for analysis. An inexperienced interviewer with a variable

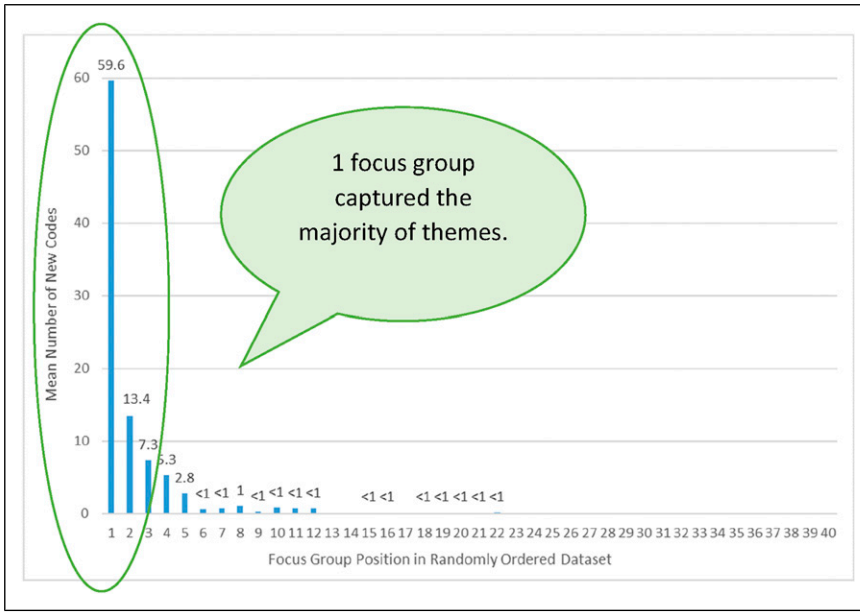

Figure 2. Achieving thematic saturation with focus groups (Source: Namey, 2017).

and very large sample could result in superficial data, providing a false sense of security and/or generating large amounts of information non-conducive to in-depth analysis (Cleary et al., 2014).

A related quality consideration is how well the discussion guide is written. One way to improve the validity of qualitative data and the utility of the insights that can be gleaned is to integrate behavioural science into the questioning approach. We know from behavioural science that people are prone to changing their responses to maintain a positive identity (social desirability bias) (Paulhus, 1984) and a consistent self-understanding (cognitive dissonance) (Festinger, 1957), or when they feel bored or find the question difficult (satisficing and acquiescence bias) (Messick, 1967; Simon, 1956). Therefore, we need to carefully frame qualitative questions to minimise response bias. Equally, the evaluation findings will be more useful to decision-makers if they lead to practical recommendations for future program design. So, we also need to design questions that test behavioural theory about the cognitive and emotional processes that sit behind responses – to explain the why, not just the what.

Even when high quality data are collected, inappropriate analytical techniques can leave insights on the table or generate misleading findings. For example, evaluators, like all humans, are prone to confirmation bias (Oswald, 2004) and positivity bias (Hoorens, 2014). Confirmation bias involves evaluators searching for evidence that confirms their prior beliefs. Positivity bias involves evaluators searching for findings

that tell a positive story and not reporting negative findings. Framework analysis helps to minimise confirmation bias and positivity bias because it forces evaluators to pre-specify how they will define success before analysing the data. However, when the goal of qualitative data analysis is to build new theory, evaluators can still enhance the rigour of emergent thematic analysis by building in time for quality assurance checks by an independent reviewer.

Skilled evaluators will demonstrate that they can apply these analytical techniques appropriately and early career evaluators should demonstrate they have a quality assurance system in place. This may be more important for achieving value for money than the number of interviews or focus groups conducted.

Conclusion

Qualitative data have a distinct role in evaluations and can provide important insights for decision-making. Qualitative findings are often centred on stakeholder experience and the how and why of a particular issue, process, situation, subculture, scene, or set of social interactions (Dworkin, 2012).

Data collection represents value for money when it provides reliable (stable across samples) and valid (true) evaluation findings. Because we are usually not trying to identify the prevalence of an issue with qualitative data, but rather why an issue occurred and how we can fix it, thematic saturation is not always important. When thematic saturation is important for answering an evaluation question, 6–7 interviews or 1–2 focus groups per cohort will capture the majority of themes. However, sometimes as few as 5 interviews or 1 focus group will achieve 90% saturation if the study population is homogenous and the evaluation question is narrowly defined.

Evaluators and those who procure evaluations should remember that sample size may be less important for generating reliable and valid qualitative findings than:

- Whether a rationale for sample selection is provided and demonstrates alignment with the evaluation question(s).
- The design of qualitative questions and the skills of the evaluator to elicit consistent, unbiased responses from evaluation participants.
- The nature of the analysis and the skills of the evaluator to draw meaningful insights.
- The thoroughness of quality assurance checks.

Therefore, it is not sufficient to assess value for money by sample size alone and consideration also must be given to the characteristics of the evaluation design and the skills of the evaluators undertaking the project.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kizzy Gandy https://orcid.org/0009-0008-3404-9423

Note

1. The bootstrap method is a resampling technique that uses the variability within a sample to estimate the sampling distribution of metrics (in this case saturation metrics) empirically. This is done by randomly resampling from the sample with replacement (i.e. an item may be selected more than once in a resample) many times in a way that mimics the original sampling scheme (Guest, Namey & Chen, 2020).

References

Better Evaluation. (2022). *Specify the key evaluation questions*. Better Evaluation. https://www.betterevaluation.org/frameworks-guides/rainbow-framework/frame/specify-key-evaluation-questions

Boddy, C. R. (2016). Sample size for qualitative research. *Qualitative Market Research: An International Journal*, *19*(4), 426–432. https://doi.org/10.1108/QMR-06-2016-0053

Cleary, M., Horsfall, J., & Hayter, M. (2014). Data collection and sampling in qualitative research: Does size matter? *Journal of Advanced Nursing*, *70*(3), 473–475. https://doi.org/10.1111/jan.12163

Dworkin, S. L. (2012). Sample size policy for qualitative studies using in-depth interviews. *Archives of Sexual Behavior*, *41*(6), 1319–1320. https://doi.org/10.1007/s10508-012-0016-6

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Guest, G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PLoS One*, *15*(5), Article e0232076. https://doi.org/10.1371/journal.pone.0232076

Guest, G., Namey, E., & McKenna, K. (2017). How many focus groups are enough? Building an evidence base for nonprobability sample sizes. *Field Methods*, *29*(1), 3–22. https://doi.org/10.1177/1525822X16639015

Guest, G., Namey, E., Taylor, J., Eley, N., & McKenna, K. (2017). Comparing focus groups and individual interviews: Findings from a randomized study. *International Journal of Social Research Methodology*, *20*(6), 693–708. https://doi.org/10.1080/13645579.2017.1281601

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science and Medicine*, *292*(1), Article 114523. https://doi.org/10.1016/j.socscimed.2021.114523

Hoorens, V. (2014). Positivity bias. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 4938–4941). Springer. https://doi.org/10.1007/978-94-007-0753-5_2219

Martínez-Mesa, J., González-Chica, D. A., Bastos, J. L., Bonamigo, R. R., & Duquia, R. P. (2014). Sample size: How many participants do I need in my research? *Anais brasileiros de dermatologia*, *89*(4), 609–615. https://doi.org/10.1590/abd1806-4841.20143705

Messick, S. (1967). The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. i–44). Aldine.

Namey, E. (2017). Riddle me this: How many interviews (or focus groups) are enough. https://researchforevidence.fhi360.org/riddle-me-this-how-many-interviews-or-focus-groups-are-enough

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 79–96). Psychology Press.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

van Rijnsoever, F. J. (2017). (I can't get No) saturation: A simulation and guidelines for sample sizes in qualitative research. *PLoS One*, *12*(7), Article e0181689. https://doi.org/10.1371/journal.pone.0181689

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, *18*(1), 148–218. https://doi.org/10.1186/s12874-018-0594-7