International Journal of Assessment Tools in Education
2020, Vol. 7, No. 1, 73–79
https://dx.doi.org/10.21449/ijate.693217
Published at http://www.ijate.net http://dergipark.org.tr/en/pub/ijate *Rapid Communications*

When interviewing: how many is enough?

William W. Cobern [1,*], **Betty AJ Adams** [1]

[1]The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

ARTICLE HISTORY

Received: 17 December 2019
Accepted: 11 February 2020

KEYWORDS

Research methodology,
Sample size,
Generalization,
Interview research,
Qualitative research,
External validity

Abstract: Researchers need to know what is an appropriate sample size for interview work, but how does one decide upon an acceptable number of people to interview? This question is not relevant to case study work where one would typically interview every member of a case, or in situations where it is both desirable and feasible to interview all target population members. However, in much of qualitative and mixed-methods research and evaluation, the researcher can only reasonably interview a subset of the target population. How big or small should that subset be? This paper provides a brief explanation of why the concept of generalization is inappropriate with respect to the findings from qualitative interviewing, what wording to use in place of generalization, and how one should decide on sample size for interviews.

1. INTRODUCTION

Researchers need to know what is an appropriate sample size for interview work, but how does one decide upon an acceptable number of people to interview? This question is not relevant to case study work where one would typically interview every member of a case, or in situations where it is both desirable and feasible to interview all target population members. However, in much of qualitative and mixed-methods research and evaluation, the researcher can only reasonably interview a subset of the target population. How big or small should that subset be?

We raise this issue because we have seen sample size, or interview numbers, questioned by both graduate students and faculty, but without much validation for their opinions. For example, a doctoral student of ours proposed to interview 20 parents of primary, middle, and high school students. The proposed 20 parents would thus be divided over three grade bands. This proposal was challenged by a few faculty and other graduate students for having too few parents in each band. These dissenters objected, argued that dividing 20 interviews across three grade bands would mean too small of an N per group, too few subjects in each band for generalization purposes. Subsequently the student and his committee decided he should focus on only one grade band, but for reasons unrelated to generalizability or N-size as voiced by the dissenters during the public presentation.

CONTACT: William W. Cobern ✉ bill.cobern@wmich.edu The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

ISSN-e: 2148-7456 /© IJATE 2020

With respect to interview work, the concept of generalization is misapplied, so on this point the student's objectors were mistaken. As it happens, efforts to misapply generalizability standards to purposive, qualitative research sampling is not uncommon among people who primarily do probabilistic, quantitative research. Still, there is a valid underlying question: what is an acceptable number for interview work? What follows is a brief explanation of why the concept of generalization is inappropriate with respect to the findings from qualitative interviewing, what wording to use in place of generalization, and how one should decide on interview number.

2. GENERALIZATION IS A STATISTICAL CONCEPT

The related concepts of generalization and sample size (N size) are from quantitative work (see for example, Teo, 2013). They have no counterparts in qualitative research including qualitative interviewing. Generalizability is a statistical concept that is often defended partially on the basis of finding a low enough resultant p-value, or probability value. If one uses the common significance level (alpha) or threshold of $p<0.05$ for statistical significance, it suggests about a 5% chance of getting this (or a more extreme) result by chance instead of as an accurate representation of a larger population. However, many conditions apply, including assumptions regarding data distribution modes, variance, and normality, for both the sample population and the larger population you might like to "generalize" about. There is rarely any certainty involved, and this is arguably even more true in education research than in medical trials or physics experiments, for instance, when comparing a control student group to a treatment student group for an instructional innovation. A resulting low p-value suggests that the null hypothesis (no difference between) is not true, but this does not necessarily mean that the treatment hypothesis is perfectly true.

Sample size expressed as an N value is related to the statistical concept of generalization through power calculations. Admittedly, researchers often neglect this calculation (typically because they are using convenience samples), but power calculations are used for estimating the N size needed to show statistically significant difference if such a difference exists.

> In plain English, statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected. If statistical power is high, the probability of making a Type II error, or concluding there is no effect when, in fact, there is one, goes down… Statistical power is affected chiefly by the size of the effect and the size of the sample used to detect it. Bigger effects are easier to detect than smaller effects, while large samples offer greater test sensitivity than small samples (Ellis, 2010).

As you can see, the ability to detect a true effect is sensitive to sample size. Hence, the ability to generalize is sensitive to sample size (Royall, 1986). However, statistical significance does not necessarily mean practical significance. A large enough sample size may allow the researcher to determine statistically that a very small difference between treatment and control conditions is significant where the difference is too small to have any practical value.

In qualitative work, such calculations do not exist and therefore the concept of generalization should not be applied to qualitative work. Nevertheless, it is good news for qualitative researchers that size isn't everything, not even in quantitative research. Indeed, years ago Cronbach offered the following advice on generalizing from quantitative data, advice insufficiently heeded by quantitative researchers:

> Instead of making generalization the ruling consideration in our research, I suggest that we reverse our priorities. An observer collecting data in one particular situation is in a position to appraise a practice or proposition in that setting, observing effects in context. In trying to describe and account for what happened, he will give attention to whatever variables were controlled, but he will give equally careful attention to uncontrolled

> conditions, to personal characteristics, and to events that occurred during the treatment and measurement. As he goes from situation to situation, his first task is to describe and interpret the effect anew in each locale, perhaps taking into account factors unique to that local of series of events (cf. Geertz, 1973, chap. 1, on "thick description"). As results accumulate, a person who seeks understanding will do his best to trace how the uncontrolled factors could have caused local departures from the modal effect. That is, generalization comes late, and the exception is taken as seriously as the rule. (Cronbach, 1975, p. 124-125)

In this quote, Cronbach refers to Clifford Geertz and his notion of "thick description" which is a notion well-known amongst qualitative researchers. The point is that even in quantitative research, the qualitative description of relevant factors is essential to the understanding of practical significance.

3. EXTERNAL VALIDITY AND QUALITATIVE INTERVIEWING

One can find discussions about sampling and generalizability in the literature on qualitative research (e.g., Gobo, 2007), but rather than speaking about generalization one should think in terms of external validity (Kukul & Ganguli, 2012). We can say that qualitative findings will be externally valid for situations similar to the one in which the study was conducted. Hence, rather than talking about how generalizable the qualitative data is, the qualitative researcher is well advised to use forms of the word "indicative" and similar words such as "suggest." The qualitative researcher should say something like "the findings of this study are indicative of what one would find in other situations given similar characteristics." Or, "this study indicates that in other situations…" Or, "this study suggests that in other situations…" Using wording such as this highlights the importance of context, which, as per Cronbach, is something that even quantitative researchers should be heeding. The qualitative researchers are saying that these findings are likely to be valid for similar situations. It is then up to consumers of the research to judge to what extent the research findings are valid for the particular circumstances of interest to that consumer. Furthermore, we do not advise that qualitative researchers use the word generalization when addressing the limitations of their work. Again, it is the language of "indication" and "suggestion" that is appropriate. The true limitation is that qualitative findings are indicative only for situations having similar characteristics.

4. SAMPLE SIZE AND THE CONCEPT OF 'SATURATION'

But we still haven't answered the question of how many to interview. The number does matter, though not for the reasons that numbers matter in quantitative work. Take for example an opinion survey where the subjects respond to items such as Likert items. The researcher needs a sample size ample enough to allow accurate estimation of how likely (probable) it is that people (of similar characteristics) will hold the opinions represented by the items. The situation is not much different with test scores. If achievement scores from treatment and control conditions are to be compared, researchers need numbers so that they can accurately estimate how likely (probable) it is that the outcome will be the same for other students (of similar characteristics). In contrast, an interview is used to determine what opinions are held by interviewees. Hence, you need to interview enough people so that you learn most if not all possible opinions (among people of similar characteristics). Of course, researchers often want to know which opinions are more popular or more frequent, but that's not the primary aim of qualitative work. Those questions are better answered quantitatively.

For qualitative interviewing there is a critical assumption: the number of unique opinions is not very large. For example, if we asked professors what they thought about working at their university there would be a limited number of opinions; from 100 professors you are not going to get 100 unique opinions. What you will find is that several opinions get repeated over and

over, which means that the researcher does not need to interview all 100 professors in order to discover all of the unique opinions in this group of people, especially not all the most common, unique opinions. Clearly, judgement is called for (see Baker & Edwards, 2012, for a variety of opinions). Here is a counter example. We were interested in how students understood a common claim about the nature of science: Scientific knowledge is durable but can change in light of new evidence or new perspectives. We particularly wanted to know how in this context students interpreted the word 'durable.' We reasoned that students could easily have more unique opinions than the number of students we could reasonably interview. Hence, we used a survey method; and the survey results validated our judgment: student opinions were many. No reasonable number of interviews would have so efficiently disclosed such a large number of opinions.

The research student we spoke of earlier, however, wanted to know what local parents thought of the new Next Generation Science Standards (NGSS) being implemented in the area schools. Knowing that parents did not have much experience with this new curriculum, he reasoned that there would be a limited number of unique opinions, and that these could be adequately identified by interviewing a subset of parents. He reasonably expected that as he went down his list of parents, a few opinions would begin reoccurring; because opinions on most topics do not run in the hundreds; they do not even run in the dozens. Unless a topic is vague, lacking focus, or poorly defined, there just are not that many distinct opinions that one could hold about most topics. The goal of qualitative interviewing is to capture most if not all of those opinions, however many opinions there are. And this is where the number of people needed for interviewing comes into question.

Clearly, the likelihood of capturing most if not all opinions increases with the number of people one interviews. The thing is, once you have captured the possible range of opinions, to whatever level of detail you seek, there is little reason to continue interviewing more people. You have reached "saturation" (Seidman, 2006). Interviewing more people will not result in more opinions because very likely there are no more opinions. The probability that a unique opinion exists is inversely related to how long it takes to find that unique opinion. But still we have to ask how many interviews are enough. One approach to deciding, and it is one that we've used, is that you don't estimate ahead of time how many people to interview. You keep interviewing until you reach a point where you stop getting unique opinions and all that you are hearing is what you have heard from previous interviewees. At that point you interview perhaps one, two, or three more for insurance; but you have reached the number you need. In a Cobern, Gibson & Underwood (1999) study, the researchers quit at 16 interviews having reached saturation.

On the other hand, oftentimes for logistical reasons, time constraints, and financial ability to pay honoraria, a researcher must decide ahead of time the maximum number of people to interview. This is the situation in which many researchers find themselves, and it calls for judgment. Researchers have to consider how many opinions on any given topic the people of interest to them might hold. Is the topic like defining 'durable' in the context of science, or asking parents their opinion of a newly implemented science curriculum? Only two opinions? Three? Three to five? Could there be 10 distinct opinions on the topic of interest? The literature can help because it can suggest what opinions might be out there, but conventional wisdom (maybe we would even say common sense) is that for most well-defined topics there are not 10 unique opinions among similar people. If we assume that there will be no more than 10 unique opinions on most of the topics we would want to ask people about, then we have to ask how many people we would need to interview to get those 10 opinions. That is the question the researcher must answer. Rather, the researcher must estimate an answer for that question. That estimation gives you the number of people you should plan to interview. Conventional wisdom suggests that the number is between 15 and 20 insofar as the topic is of limited scope. It was a

good bet that the high schoolers' parents our doctoral student was interested in would have fewer than 20 unique opinions about NGSS, and that those opinions might or might not be equally common. By the 20th interview, he could expect to have reached saturation – and he did (Channell, 2019) (see Appendix for how this approach might be worded for a research proposal.)

Our point is that for qualitative interviewing, the number of people one plans to interview is not the first question that needs to be answered. For our graduate student, the important question was, how likely are parents of students, across the three grade bands, to have such differing opinions that the domain of unique opinions across the three grade bands exceeds the number of unique opinions in any one grade band. If it can be argued that grade band is unimportant, then his original plan was fine. On the other hand, if different grade bands are likely to result in different opinions, then six or seven interviews per grade band would not likely be enough to reach saturation per grade band, and too few would probably be too risky.

5. CONCLUSION

All research requires judgement. It does not matter whether the research is quantitative or qualitative; judgment is required. Not even a power calculation can be run without judgment, because the input values are not self-evident. A good quantitative researcher describes the situation in which the research takes place and defends value judgments and assumptions. A qualitative researcher does the same. Deciding on how many subjects to interview is a value judgment and requires an explanation. We knew from other research that students very likely had a poor understanding of what it meant that scientific knowledge is 'durable.' Hence, we could reasonably expect that a large number of students would hold a number of unique opinions almost as large, thus making an interview approach not only impractical but nigh impossible. On the other hand, NGSS is a new curriculum in our area and parents simply had not had much time to form many opinions. Moreover, the focus on NGSS was specific to the science classrooms where the parents' children attended. An interview approach was reasonable. The choice between administering quantifiable surveys or conducting qualitative interviews does not usually require elaborate explanation. However, for interview work, we advise explaining the general basis for sample size, and also whether or not informational redundancy or saturation was achieved.

Finally, we urge qualitative interviewers to exchange the rhetoric of generalizing for the rhetoric of external validity. Some research is designed to simply provide specific and actionable information about the sample population. More often, consumers of research want to know whether qualitative findings are applicable to their own situation of interest or indicative of what might be the case in a different but similar situation. This is the judgment that consumers of research have to make and that they can only make if the original researchers adequately describe the context in which the research was conducted. If you're going to interview parents about their children's education, we need descriptive information about the parents and about the schools that their children attend. Only then can the consumer judge whether or not aspects of the findings are likely to be valid elsewhere, that is, judge to what extent the findings have external validity. Generalization, however, is a term best left for quantitative, probability-based research where, even then, generalizing applies to adequately similar situations or populations. Qualitative findings can be usefully indicative of what one might find in similar situations and contexts, and also of how different aspects/elements studied may relate to one another.

Acknowledgements

Not applicable.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

William W. Cobern http://orcid.org/0000-0002-0219-203X
Betty AJ Adams http://orcid.org/0000-0002-8554-8002

6. REFERENCES

Baker, S. E., & 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 Review Paper. Retrieved December 28, 2019 from http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf

Channell, A. C. (2019). Teacher and Parent Perspectives on Alignment to The Next Generation Science Standards Following Teacher Professional Development. (PhD), Western Michigan University, Kalamazoo, MI.

Cobern, W. W., Gibson, A. T., & Underwood, S. A. (1999). Conceptualizations of Nature: An Interpretive Study of 16 Ninth Graders' Everyday Thinking. *Journal of Research in Science Teaching*, *36*(5), 541-564. DOI.org/10.1002/(SICI)1098-2736(199905)36:5<541:AID-TEA3>3.0.CO;2-1

Cronbach, L. J. (1975). Beyond the Two Disciplines of Scientific Psychology. *American Psychologist*, *30*(2), 116-127. DOI:10.1037/h0076829

Ellis, P. D. (2010). Effect Size FAQs. Retrieved December 20, 2019 from https://effectsizefaq.com/about/

Gobo, G. (2007). Sampling, representativeness and generalizability. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative Research Practice*. SAGE Publications: Thousand Oaks, CA, p. 405-426. ISBN-13: 978-0761947769.

Kukull, W. A., & Ganguli, M. (2012). Generalizability: the Trees, the Forest, and the Low-Hanging Fruit. *Neurology*, *78*(23), 1886-1891. DOI:10.1212/WNL.0b013e318258f812.

Royall, R. M. (1986). The Effect of Sample Size on the Meaning of Significance Tests. *The American Statistician*, *40*(4), 313-315. DOI:10.2307/2684616

Seidman, I. E. (2006). *Interviewing as Qualitative Research: A Guide for Researchers in Education and The Social Sciences, 3rd Edition*. Teachers College Press: Columbia University, New York. ISBN-13: 978-0807746660.

Teo, T. (2013) (Ed.). *Handbook of Quantitative Methods for Educational Research*. Sense Publishers: Rotterdam, The Netherlands. ISBN: 978-94-6209-404-8.

7. APPENDIX

Here is an example of how this approach might be worded for a research proposal. For this example, we are indebted to our colleague Dr Brandy Pleasants.

> Based on research with a similarly homogenous group it seems that about 10 participants is sufficient to cover all reasonable responses I might get. I therefore plan to interview no less than 10 participants, with a goal of 15 (even if saturation is reached); however, I also plan to continue interviewing if at 10 I still seeing variation in the data, continuing until I reached saturation.