Beyond the Guise of Saturation: Rigor and Qualitative Interview Data

Kori A. LaDonna, PhD
Anthony R. Artino Jr, PhD
Dorene F. Balmer, PhD

Health professions education researchers, including those who study graduate medical education (GME), are building an evidence base to guide educational practice. Over the last 2 decades, qualitative researchers have generated a plethora of empirical findings. However, what are the features of good qualitative evidence? In our teaching and mentorship roles, we are increasingly asked to counsel colleagues who are tentatively dipping their toes into qualitative waters to ask research questions that cannot be answered using quantitative methods. In our reviewer and editor roles, we have noticed that authors sometimes make substantive claims based on qualitative interview data that have concerning limitations. Moreover, in our researcher roles, we must carefully defend the soundness of our qualitative findings, not only because doing so is a good research practice, but also because we anticipate that some reviewers may erroneously apply quantitative criteria to evaluate qualitative methods.

These experiences highlight the need for further clarification about evaluating the "evidentiary value" of qualitative findings for informing pedagogy and improving practice. We argue that evidentiary value depends not only on the rigor of the research process and the richness of data generated during interviews, but also on how clearly and effectively investigators report their findings and demonstrate their contributions to GME. However, the diversity of expertise in GME means that the value of qualitative research is often in the eye of the beholder. In this editorial, we discuss the features of high-quality evidence obtained through interviews and provide guidance to help GME researchers, reviewers, and readers recognize valuable qualitative evidence when they see it.

Rigor and Saturation

In health professions education, qualitative researchers explore *how* and *why* questions, such as "How do faculty members navigate underperformance or failure?"[1] or "Why do some medical students maintain a career interest in pediatrics while others do not?"[2] To answer such questions, the qualitative research process must be appropriately robust to produce findings that are transferable rather than generalizable,[3] which means that they provoke thought, raise questions, and inform or change practice in settings beyond the research context. To do this, findings do not need to be valid, reliable, or representative, but they do need to be credible, resonant, and rich.[3–6] Given the subjectiveness of these criteria, how do we evaluate the rigor of qualitative research that uses one-on-one interview methods?

Rigor is often assumed to hinge largely on *saturation*, which is typically understood as the point in data collection where interviews are either no longer generating new information or when researchers determine that they have "heard it all."[7] While this idea seems simple enough, considerable confusion about what saturation means makes it difficult to determine when (or if) it is reached. Indeed, a systematic analysis of qualitative interview-based studies demonstrated that authors variably and inadequately described indices of saturation and often focused on participant numbers to try to convince reviewers (and perhaps themselves) that they have recruited a large enough sample to substantiate their claims.[8] Consequently, many GME researchers make statements like "we reached saturation after the ninth resident was interviewed" without either describing what saturation means for their study or providing evidence to support the claim that their data were actually saturated.[8]

Shifting From Saturation to Sufficiency

A qualitative dataset should be comprehensive enough (depth) to both identify recurrent thematic patterns and to account for discrepant examples (breadth).[7] In other words, saturation depends on more than the number of participants. We caution reviewers that appraisals of quality focused primarily on sample size may be a guise for data that do not

DOI: http://dx.doi.org/10.4300/JGME-D-21-00752.1

meet these criteria. In fact, a recent international study of research ethics found that 11% of researchers admitted to knowingly using terms like saturation improperly, making it among the most common questionable research practices in health professions education.[9]

To further complicate matters, some qualitative researchers have begun to question whether reaching saturation is even possible.[10–13] Instead, many qualitative researchers have shifted to describing quality findings as *sufficient,* recognizing that sufficiency depends on both the rigor of the analytical process (analytical sufficiency) and the richness of the data it generates (data sufficiency). Unlike saturation, which likens a dataset to a sponge with an objective saturation point, the notion of sufficiency suggests that—within a research paradigm that acknowledges both the uniqueness of human experience and the socially constructed nature of data—researchers can metaphorically wring out their dataset, continuously dipping into a well of new understanding by iteratively revising interview guides, sampling new participants, and engaging in multiple rounds of data generation and analysis. But research studies cannot go on forever. Without power analyses or sample size calculations to rely on, how can researchers convincingly demonstrate not that they have "heard it all,"[7] but that they have heard *enough*?

Evaluating the Sufficiency of Qualitative Findings

Given the limitations of the saturation concept, the notion of *information power*[14] may provide a better gauge for evaluating sufficiency. Using information power to determine whether qualitative findings are sufficient depends on examining them alongside the aims of the study, the specificity of the sample, the use of theory, the strategy for analysis, and the quality of the interviews.

Qualitative researchers use a multitude of methodological approaches that draw on various analytical strategies to examine a phenomenon from a distinct vantage point. Some methodologies are designed to produce an in-depth analysis of a few individual accounts, whereas other methodologies require a larger sample to analyze a phenomenon from multiple points of view.[14] Moreover, a narrower study aim with a targeted group of potential participants may allow for data sufficiency to be achieved with a leaner sample size. To illustrate, consider that a study exploring how child abuse fellows in Texas and New Mexico manage their first case of suspected rape by human smugglers may need fewer participants than a study with the much broader aim of examining how pediatrics fellows across North America manage their emotions when reporting child abuse.

Requirements for sufficiency also depend on whether the researcher's intention is to describe a phenomenon or to generate theory. For example, a descriptive qualitative study[15] of first-year residents engaging with virtual learning will likely require both a smaller sample of interviews and less intensive analytical work than a constructivist grounded theory (CGT)[16,17] exploration of adaptations to virtual learning. In CGT, robust theorizing often relies on 20 or more in-depth interviews[18] and multiple rounds of increasingly interpretive coding.[16,17] Indeed, studies using theory *a priori* to examine a phenomenon through a specific research lens are at different starting points for reaching sufficiency than studies seeking to build theory inductively. Consequently, a study using self-determination theory[19] to frame data generation and analysis will likely reach sufficiency with fewer interviews and less interpretive labor than a study aimed at generating theory about residents' motivation to engage in learning outside the formal curriculum.

The information power model dispels the myth that bigger samples equal better data. Thus, when evaluating sufficiency, interview quality matters more than quantity. To generate rich data, interviews must be conversational, focused on the research topic, and peppered with strategic follow-up questions and prompts for illustrative examples. Interviewer skill is paramount. Interviewers need to develop rapport with participants, invite thoughtful reflection, and adapt the interview guide to allow for research questions to expand or shift direction depending on participants' in-the-moment responses and the evolving analysis. While we hesitate to quantify qualitative rigor, we suggest that interview length may be a more useful indicator of information power than sample size. While this guidance is not foolproof and should not be followed prescriptively, 6 in-depth interviews with open-ended questions lasting an hour or more will likely yield richer data than twenty 10-minute interviews that elicit only surface-level responses. Of course, the true test of sufficiency is whether interview data are not only rich but also contribute new or thought-provoking insights into a GME concept, practice, or problem.

Clearly Conveying Evidentiary Value

We warn researchers that ineffective scholarly writing can make even the most powerful qualitative findings appear unconvincing. While information power is useful for appraising or justifying the sufficiency of a qualitative sample, the evidentiary value of qualitative

TABLE
Guiding Questions When Evaluating Evidentiary Value of Qualitative Interviewing Findings

Section	Questions
Introduction	▪ Have the authors adequately mapped the literature to identify why their study matters and what it contributes?[20,21] ▪ What is the study aim and is it broad or narrow?
Methods	▪ Does the study team have appropriate qualitative expertise? Do reviewers have appropriate expertise to evaluate this work? ▪ Was an adequate rationale provided for methodological choices, including use of one-on-one interviews? ▪ Who was sampled and why? Does the sample fit with the study aim? Does the sample size have a theory to lean on? ▪ Since interviews are conversational and interview guides are expected to evolve, asking interviewers to append an interview guide to their manuscript may be challenging. Instead, do the researchers provide a sense of the interview questions they asked? Is it clear from the study aim why researchers asked certain interview questions? ▪ Is interview length noted? ▪ Is the iterative data generation and analysis process clearly described? Have discrepant cases or examples been considered?
Results	▪ Is it reasonable to assume that the interview questions generated the reported data? ▪ Do quotes from interviews substantiate claims? While it is unnecessary to provide a numerical count, do the authors indicate whether the presented findings were shared by all, by some, or by a few participants? Are multiple participant accounts represented, or do results lean on a handful of interviews? ▪ Do the quotes provide cohesive, rich accounts of a phenomenon? Are the findings presented as a list of themes that the reader needs to piece together, or have researchers developed a coherent analysis? ▪ Do the results make sense? Do they seem plausible?
Discussion	▪ Have the authors considered their findings alongside the relevant literature, and do they clearly and convincingly describe what their qualitative findings add to the health professions education evidence base? ▪ Have the authors thoughtfully considered the strengths and limitations of their research? ▪ Have the authors demonstrated that findings are transferable beyond their research context? Do they describe why study limitations interfere with transferability (not just that they do)? Have the reviewers and readers learned anything?

findings depends on more than rich data and rigorous analysis. It requires good writing. When drafting research for publication, the onus is on the authors to make their research procedures and decision-making processes transparent and convincing.[8,22] Researchers need to clearly and compellingly convey not only *why* a dataset is sufficient, but also *how* data were interpreted and *what* they contribute to GME. Enumerating a list of disparate themes, rather than demonstrating how themes connect to generate new understanding, will likely fail to convince reviewers and readers that the findings are meaningful. In turn, reviewers and readers must be mindful that strong manuscripts may wilt under the inappropriate application of quantitative criteria that fail to capture the nuances of rigorous qualitative research.

Boosting the qualitative evidence base in GME depends on both demonstrating sufficiency and evaluating it appropriately. In the TABLE we provide a set of guiding questions to consider when evaluating or reporting the evidentiary value of qualitative interview findings.

Summary

We urge GME researchers, reviewers, and readers to move beyond the guise of saturation when evaluating qualitative findings obtained from interviews. In this editorial, we provide guidance to help qualitative novices develop a scholarly language to articulate—and in some cases, check—their gut sense about the evidentiary value of qualitative interview data. However, given the complexities of qualitative research, our guidance is written in sand, not stone. We hope that the list of guiding questions (TABLE) and key references (BOX) will promote deeper reflection and learning around these important qualitative issues. We encourage GME researchers, reviewers, and readers to thoughtfully use concepts like richness, rigor, sufficiency, and information power, and to seek advice from qualitative research experts when in doubt.

BOX Additional Resources

- Frambach JM, van der Vleuten CPM, Durning SJ. AM Last Page: Quality criteria in qualitative and quantitative research. *Acad Med*. 2013;88(4):552. doi:10.1097/ACM.0b013e31828abf7f
- Lingard L, Watling C. *Story, Not Study: 30 Brief Lessons to Inspire Health Researchers as Writers*. Cham, Switzerland: Springer; 2021.
- Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. *Qual Health Res*. 2016;26(13):1753–1760. doi:10.1177/1049732315617444.
- O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. *Acad Med*. 2014;89(9):1245–1251. doi:10.1097/ACM.0000000000000388
- Paradis E, O'Brien B, Nimmon L, Bandiera G, Martimianakis MA. Design: selection of data collection methods. *J Grad Med Educ*. 2016;8(2):263–264. doi:10.4300/JGME-D-16-00098.1
- Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. *Med Educ*. 2016;51(1):40–50. doi:10.1111/medu.13124
- Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Med Res Methodol*. 2018;18:148. doi:10.1186/s12874-018-0594-7
- Wright S, O'Brien BC, Nimmon L, Law M, Mylopoulos M. Research design considerations. *J Grad Med Educ*. 2016;8(1):97–98. doi:10.4300/JGME-D-15-00566.1

References

1. LaDonna KA, Ginsburg S, Watling C. Shifting and sharing: academic physicians' strategies for navigating underperformance and failure. *Acad Med*. 2018;93(11):1713–1718. doi:10.1097/ACM.0000000000002292
2. Balmer DF, Gottlieb-Smith RJ, Hobday PM, et al. Pediatric career choice: insights from a novel, medical school pathway program. *Acad Pediatr*. 2020;20(1):97–103. doi:10.1016/j.acap.2019.07.013
3. Lincoln YS, Guba EG. *Naturalistic Inquiry*. Thousand Oaks, CA: Sage Publications; 1985:416.
4. Balmer DF, Rama JA, Athina Tina Martimianakis M, Stenfors-Hayes T. Using data from program evaluations for qualitative research. *J Grad Med Educ*. 2016;8(5):773–774. doi:10.4300/JGME-D-16-00540.1
5. LaDonna KA, Taylor T, Lingard L. Why open-ended survey questions are unlikely to support rigorous qualitative insights. *Acad Med*. 2018;93(3):347–349. doi:10.1097/ACM.0000000000002088
6. Tracy SJ, Hinrichs MM. Big tent criteria for qualitative quality. *Int Encyclopedia Commu Res Method*. 2017:1–10. doi:10.1002/9781118901731.iecrm0016
7. Morse JM. "Data were saturated . . ." *Qual Health Res*. 2015;25(5):587–588. doi:10.1177/1049732315576699
8. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Med Res Methodol*. 2018;18(1):148. doi:10.1186/s12874-018-0594-7
9. Artino AR Jr, Driessen EW, Maggio LA. Ethical shades of gray: international frequency of scientific misconduct and questionable research practices in health professions education. *Acad Med*. 2019;94(1):76–84. doi:10.1097/ACM.0000000000002412
10. O'Reilly M, Parker N. "Unsatisfactory Saturation": a critical exploration of the notion of saturated sample sizes in qualitative research. *Qual Res*. 2013;13(2):190–197. doi:10.1177/1468794112446106
11. Braun V, Clarke V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qual Res Sport Exercise Health*. 2019;13(2):201–216. doi:10.1080/2159676X.2019.1704846
12. Nelson J. Using conceptual depth criteria: addressing the challenge of reaching saturation in qualitative research. *Qual Res*. 2017;17(5):554–570. doi:10.1177/1468794116679873
13. Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. *Med Educ*. 2017;51(1):40–50. doi:10.1111/medu.13124
14. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. *Qual Health Res*. 2016;26(13):1753–1760. doi:10.1177/1049732315617444
15. Sandelowski M. Whatever happened to qualitative description? *Res Nurs Health*. 2000;23(4):334–340. doi:10.1002/1098-240x(200008)23:4<334::aid-nur9>3.0.co;2-g
16. Charmaz K. *Constructing Grounded Theory*. 2nd ed. Thousand Oaks, CA: Sage Publications; 2014.
17. Watling C, Cristancho S, Wright S, Varpio L. Necessary groundwork: planning a strong grounded theory study. *J Grad Med Educ*. 2017;9(1):129–130. doi:10.4300/JGME-D-16-00693.1
18. Thomson SB. Sample size and grounded theory. *J Admin Governance*. 2010;5(1):45–52.
19. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol*. 2000;55(1):68–78. doi:10.1037//0003-066x.55.1.68
20. Lingard L. Joining a conversation: the problem/gap/hook heuristic. *Perspect Med Educ*. 2015;4(5):252–253. doi:10.1007/s40037-015-0211-y

21. Lingard L. Writing an effective literature review: part I: mapping the gap. *Perspect Med Educ*. 2018;7(1):47–49. doi:10.1007/s40037-017-0401-x
22. Lingard L, Watling C. *Story, Not Study: 30 Brief Lessons to Inspire Health Researchers as Writers*. Cham, Switzerland: Springer; 2021.

Kori A. LaDonna, PhD, is Assistant Professor, Department of Innovation in Medical Education and Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada; **Anthony R. Artino Jr, PhD,** is Professor and Associate Dean for Evaluation and Educational Research, The George Washington University School of Medicine and Health Sciences, and Deputy Editor, *Journal of Graduate Medical Education*; **Dorene F. Balmer, PhD,** is Associate Professor, Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania.

The authors would like to thank Dr. Renate Kahlke, Dr. Roy Khalife, and Mr. Leif-Erik Aune for their thoughtful feedback on earlier iterations of this editorial.

Corresponding author: Anthony R. Artino Jr, PhD, The George Washington University School of Medicine and Health Sciences, aartino@email.gwu.edu, Twitter @mededdoc