Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry

Qualitative Health Research
2015, Vol. 25(9) 1212–1222

DOI: 10.1177/1049732315588501
qhr.sagepub.com
SAGE

Janice M. Morse[1]

Abstract

Criteria for determining the trustworthiness of qualitative research were introduced by Guba and Lincoln in the 1980s when they replaced terminology for achieving rigor, reliability, validity, and generalizability with dependability, credibility, and transferability. Strategies for achieving trustworthiness were also introduced. This landmark contribution to qualitative research remains in use today, with only minor modifications in format. Despite the significance of this contribution over the past four decades, the strategies recommended to achieve trustworthiness have not been critically examined. Recommendations for where, why, and how to use these strategies have not been developed, and how well they achieve their intended goal has not been examined. We do not know, for example, what impact these strategies have on the completed research. In this article, I critique these strategies. I recommend that qualitative researchers return to the terminology of social sciences, using rigor, reliability, validity, and generalizability. I then make recommendations for the appropriate use of the strategies recommended to achieve rigor: prolonged engagement, persistent observation, and thick, rich description; inter-rater reliability, negative case analysis; peer review or debriefing; clarifying researcher bias; member checking; external audits; and triangulation.

Keywords

qualitative; trustworthiness; rigor; reliability; validity; generalizability; dependability; credibility; transferability

In the 1980s, Guba and Lincoln transformed the nature of qualitative inquiry by developing criteria to ensure rigor (which they termed *trustworthiness*) during qualitative inquiry, to evaluate the credibility, transferability, dependability, and the trustworthiness of the completed product (Guba, 1981; Guba & Lincoln, 1985, 1989). These criteria included strategies that, in various forms, have been used extensively for ensuring rigor for four decades, yet, with few exceptions (Krefting, 1991; O'Neill, 1995; Tuckett, 2005), have not been questioned. They have been used as the criteria to ensure rigor during the conduct of the research, as checklists for the evaluation of the research, and as indicators of the worthiness of the research (Morse, in press-a, in press-b). All textbooks on qualitative methods contain a description of these strategies to be used in inquiry; funding agencies use them to provide an assurance of quality, and journal reviewers use them for the evaluation of the reliability and the validity of the completed research.

However, despite these strategies being universally perceived as essential in qualitative inquiry, their use has not been carefully examined. We have neither critiqued their use, nor determined how they actually affect research design and the quality of inquiry. We have not even determined which strategies are appropriate for use in each type of inquiry, or for each type of data collection strategy. We have not determined how they actually ensure the quality of inquiry and how they change the final product of our research.

What are these strategies? Guba and Lincoln (1989, pp. 301–329) identified the overall goal of *trustworthiness*, consisting of credibility, transferability, dependability, and confirmability, to be respectively equivalent to quantitative criteria of internal validity, external validity, reliability, and objectivity. These criteria have remained the same over time (see, for instance, Creswell, 2012), but the strategies for attaining each have changed over time. The strategies proposed by Guba and Lincoln (1989) were as follows:

> *Credibility (i.e., internal validity)*: Prolonged engagement, persistent observation, triangulation, peer debriefing, negative case analysis, referential adequacy, and member checks (process and terminal).
> *Transferability (external validity, or generalizability)*: Thick description is essential for "someone interested"

[1]University of Utah, Salt Lake City, Utah, USA

Corresponding Author:
Janice M. Morse, University of Utah, 2151 E 900 S, Salt Lake City, UT 84108, USA.
Email: vshannonqhr@gmail.com

(p. 316) to transfer the original findings to another context, or individuals.

Dependability (i.e., reliability): Attainable through credibility, the use of "overlapping methods" (triangulation), "stepwise replication" (splitting data and duplicating the analysis), and use of an "inquiry audit" (p. 317) or audit trail.

Confirmability (Objectivity): Using strategies of triangulation and the audit trail.

The final step, addressing all of the above criteria, is the *reflexive journal.*

However, how these strategies are operationalized has developed over time, and it is not clear whether Guba and Lincoln expected all to be practiced for every project. Creswell (2012) noted that "qualitative researchers should engage in at least two of them in any given study" (p. 253), but he did not tell us which two, or why, when, and how to use each strategy. Furthermore, we do not know whether and how the use of these strategies, as presently operationalized during inquiry, affects the process of inquiry. Do they enhance the processes of conceptualization for the investigators? Inhibit inquiry? Do they *work*—making the inquiry more rigorous—or do they make no difference to the developing research? *Do they actually make our research valid, reliable and trustworthy?* These questions suggest it is time to reopen the discussion on the determination of trustworthiness in qualitative inquiry: We must critique the use of recommended strategies and reexamine our discipline.

Components of Rigor

While Guba and Lincoln's (1989) goal was trustworthiness, elsewhere (Morse, in press-b), I have suggested that it is time we return to the terminology of mainstream social science, using *rigor* (rather than trustworthiness), and replacing dependability, credibility, and transferability with the more generally used *reliability, validity*, and *generalizability*.

Validity (or internal validity) is usually defined as, the "degree to which inferences made in a study are accurate and well-founded" (Polit & Beck, 2012, p. 745). Miller (2008b) further defined it as "the 'goodness' or 'soundness' of a study" (p. 909). In qualitative inquiry, this is usually "operationalized" by how well the research represents the actual phenomenon. Is the resemblance accurate? Does the description of the essence actually match the essence of the phenomenon? Ask: Can the description be recognized by others who have had the experience, or appreciated by those who have not had the experience? Is the description detailed? Decontextualized? Abstracted? Expressed logically?

Reliability is "broadly described as the dependability, consistency, and/or repeatability of a project's data collection, interpretation, and/or analysis" (Miller, 2008a, p. 745). Basically, it is the ability to obtain the same results if the study were to be repeated.

Generalizability (or external validity) is "extending the research results, conclusions, or other accounts that are based on the study of particular individuals, setting, times or institutions, to other individuals, setting, times or institutions than those directly studied" (Maxwell & Chmiel, 2014; Polit & Beck, 2012). In qualitative inquiry, the application of the findings to another situation or population is achieved through decontextualization and abstraction of emerging concepts and theory. Although Guba and Lincoln (1989) suggested that this be the prerogative of a third party, I consider such recommendations also to be the prerogative of the original investigator.

In this article, I will argue that in qualitative inquiry, validity and reliability are often intertwined, with reliability attainment inherently integrated as processes of verification in the attainment of validity, hence, agreeing with Guba (1981) that validity may be supported internally by reliability. In addition, I will discuss that some strategies are particular to certain types of interviews or sampling strategies (selection and segmentation), and that some strategies, as presently practiced, may even invalidate the research. Finally, I will make a case for the increasing use of generalizability in qualitative research.

Achieving Rigor

Both criteria of reliability and validity are intended to make qualitative research rigorous (formerly referred to as trustworthy). Through particular representation, abstraction, and theory development, validity enables qualitative theories to be generalizable and useful when recontextualized and applied to other settings. Reliability makes replication possible, although qualitative researchers themselves recognize induction is difficult (or even impossible) to maintain with replication (Morse, 2008a), so the process of replication itself destroys induction. Therefore, replication of a project is unnecessary and undesirable in qualitative inquiry.

Nevertheless, *rigor* as a concept is an important goal, and rigor is the concern of external evaluators who ultimately determine the worth of qualitative research. Evaluators' concerns primarily arise from the intimate relationship between the researchers and their data: The unstructured processes of obtaining data within verbal interaction or observation, the interpretative nature of analysis, and the subjective nature of data itself are perceived as threats to validity. Furthermore, qualitative inquiry does not demand the checks and balances of

journalism (Morse, in press-a), is perceived as threatened by the imagination and creativity of interpretation and the threat of fiction, and is not intended to follow the stringent demands of randomization.[1] These features provide an opportunity for those who do not appreciate the internal methodological mechanisms, the checks, and balances of qualitative analytic processes. They demand demonstrable checks and balances as "proof" of rigorous inquiry. Thus, the strategies for demonstrating reliability and validity identified by Guba and Lincoln (1985) have become standards for attaining rigor in qualitative inquiry. We do not have reliability coefficients or other statistical tests of quantitative inquiry to offer as evidence. Presently, in qualitative inquiry, provided some of the strategies are named in the final report by the investigator as having been attended to in the process of inquiry, our research is considered "trustworthy." However, the appropriateness of these strategies has never been evaluated, nor do we know how they facilitate or hinder the actual practice of qualitative inquiry. Therefore, we will now explore the appropriateness of the recommended strategies for ensuring rigor.

Exploring Strategies to Determine Validity

The strategies for ensuring validity are prolonged engagement, persistent observation, and thick, rich description; negative case analysis; peer review or debriefing; clarifying researcher bias; member checking; external audits; and triangulation. Each of these will be discussed below.

Prolonged Engagement, Persistent Observation, and Thick, Rich Description

These three strategies are highly interdependent: Prolonged engagement and persistent observation are both necessary for producing thick, rich data. The assumption underlying these criteria is that spending more time on data collection in a particular setting provides time for trust to be established with participants. With increased trust (and intimacy), you will get better, richer data. More will be revealed, and therefore, data will be more valid. Thus, studies that use observation, and require contact with participants, must follow this practice, called "getting in" by ethnographers, or the first stage of fieldwork (Agar, 1996). However, with unstructured interviews, provided the participant has had time to assess the interviewer before starting the interview, in the process of relating one's experience, the participant internally focuses and provides rich data (Corbin & Morse, 2003). Time to get to know the interviewer is now considered necessary for unstructured interview research.

Nevertheless, obtaining thick and rich data is more than simply obtaining good data from one participant. Thick and rich data refer to the entire data set, so data quality is also associated with the number of interviews and/or participants. To have a thick and rich data set, the researcher must attend to sample size and appropriateness (i.e., selecting an appropriate sample) of the data.

What about sample size? Recently, there has been an extraordinary debate in qualitative research, with researchers trying to calculate sample size á priori (see, for instance, Guest, Bunce, & Johnson, 2006). Yet, we know that the size of the sample depends on the nature of the phenomenon, its concrete versus subjective nature, the amount of complexity and scope of the phenomenon, and of course, how much is already known about the topic. Into this mix, consider the quality of the interview (the interviewer's techniques and the quality of the participant's report), the nature of the research topic, the type of interview, and the researcher's analytic skill. The short answer is that trying to predetermine sample size is a futile task. But, we do know that if the sample is too small, data are not saturated, the results obtained are superficial and obvious, and cherry picking (Morse, 2010) occurs in the data. Data will be difficult to analyze, because the categories will not be evident, and theorizing will be difficult if not impossible. There may be limited variation, with the findings limited in scope, and the researcher will be uncertain about the findings. In the worst case, if the sample is small and perhaps also using semi-structured interviews, with these restricted data, the researcher will have nothing of interest to write up. The results will be predictable and easily predetermined. Thus, the lack of an adequate sample is a validity issue,[2] because the research lacks variation and depth, and can neither provide detailed understanding nor adequately represent the phenomenon.

What about sampling appropriateness? When commencing qualitative data collection, the researcher may know little about the topic. Data collection must begin somewhere, and at the beginning, data collection may be hit and miss. The researcher may use a convenience sample, interviewing everyone who volunteers, or use a quota sample, interviewing a certain number of people with particular positions of interest. Nevertheless, once the researcher's analysis of the interviews begins to provide some understanding, and categories or themes begin to develop, sampling strategy changes. As patterns and conjectures emerge, the sampling approach changes from a convenience sample to theoretical sampling. That is, rather than interviewing those who are available or who volunteer, the researcher seeks information from those most likely to know the information required to verify or to move understanding forward. Data collected may come from their own experiences or may be shadowed

data (Morse, 2001b) that comes from beyond their own experiences. Shadowed data is information about what they know more generally about the topic and the behavior of others. Thus, data collected are always about more than the person's own experiences; it is also about the experiences of those they know. In this way, qualitative studies always include information far beyond the "numbers of participants" that are listed in the demographic tables (Morse, 2008b).

Negative Case Analysis

Negative cases often provide the key to understanding the norm, or the most commonly occurring cases. They are not ignored, or as outliers in quantitative research, discarded. Rather, they are analyzed as carefully as the more frequently occurring cases, and as with all data, additional cases are sought and compared. Thus, data from negative cases must also be saturated.

What do I mean by "the key to understanding the norm"? Comparison of the negative cases with the more commonly occurring cases will reveal important differences, and it is the developing understanding of these differences that is often critical to understanding the process as a whole. Therefore, this is a critical analytic strategy for the development of validity.

Peer Review or Debriefing

Peer review or debriefing is intended to prevent bias and aid conceptual development of the study. The opportunity to present one's findings is a strategy that enables conceptualization of the theory, in particular for new investigators. It assists new researchers to synthesize and to see patterns in their data—sometimes by the questions asked by their peers, and sometimes even by listening to their own voice. As such, it assists with the development of internal validity.

However, I am not certain about "peer review" as something that facilitates validity, nor as a strategy for "debriefing." Elsewhere, I have written that validity is not something that is awarded by committee consensus (Morse, 1998), but is something that must be analytically earned. There is documentation of problems occurring when key informants and editors have demanded that their theoretical ideas be added to ethnographies ("Who's the anthropologist here?" see Nussbaum, 1998). As it is the researcher who is the closest to these data, responsible for the analysis, and the most familiar with the research agenda and relevant theories, the researcher must be responsible for the research outcome. It is recommended that the researcher listens to alternative points of view, but takes final responsibility for the results, and its implications and applications.

Development of a Coding System

Coding systems must be used cautiously with research using unstructured interview research. With unstructured interview research, it is not possible to develop a detailed priori coding system at the beginning of the study, as you are conducting this study because you know little about the phenomenon, and therefore, guessing what these codes might be may destroy validity. However, developing a coding system after the interviews have commenced will also not be possible, because the interviews change as the study progresses. Interviews using the same codes as used for interviews early in the study will not be suitable for coding more specific interviews conducted later in the study.[3]

However, it is desirable to establish a coding system for research using semi-structured interviews. These interviews are used when the researcher has more information about the topic, and the question stem format keeps the interviews relatively standardized. Obtaining the coding system to ensure that the meaning of the analysis is the same between coders enhances validity and certainty of the findings.

Clarifying Researcher Bias

There are two sources of researcher bias, and the investigator must be aware of each. First, there is the "pink elephant" bias (Morse & Mitcham, 2002; Spiers, in press). This is the tendency for the researcher to see what is anticipated. However, research may also be value-laden; if we expect a situation to have particular characteristics (e.g., anti-feminist) or use a value-laden theory, these features may be unfairly emphasized in data during our analysis. Investigators are always advised to enter a setting with a "neutral stance" (Popper, 1963/1965). The responses to this problem are difficult, as research, by definition, is always topic/question focused (Morse & Mitcham, 2002). Furthermore, if the researcher uses stepwise verification during data gathering, data will correct themselves during the processes of collection and analysis and strategies of verification will provide a system of checks and balances (Meadows & Morse, 2001; Morse, Barrett, Mayan, Olson, & Spiers, 2002).

Another source of bias is inherent in qualitative methods. Because we do not use a random sample, and because our samples are small and our data cumbersome, we select relatively small and excellent examples of what we are studying. Such sampling patterns are essential for the understanding of the concepts of interest. For example, in my examination of comforting, I elected to study comfort where patients were most in need of comfort—those in agony—so I decided to explore comfort in the trauma center (Morse, in press-c). This enabled me to develop an

understanding of comfort, so that when I explored comfort in a less extreme situation, I could distinguish the concept from the surrounding conceptual "noise." This principle of sampling is essential for validity, but the researcher must remember that the initial investigations are with the best (or most pure) examples of the behavior, and the investigator risks invalidity when conducting the research in a more "average" situation, where the "sheep" may not be as easily separated from the "goats."

The third type of bias occurs unconsciously in the research design. The questions may be biased; a comparative sample designed with non-equivalent samples, or non-equivalent experiences or interventions. For instance, a researcher may compare perceived quality of care offered by male and female nurses (a gender-biased question), yet disregard the gender of the patient in the research design. In these cases, the onus is on the researcher to be strongly vigilant about such comparisons and conclusions, maintaining an inductive perspective and verifying all data during processes of data gathering.

Member Checking

Member checking refers to giving the transcribed interview (or the completed analysis) back to the participant to obtain additional information or to correct data. It is not clear why one should provide the participant with such an opportunity to change his or her mind; it is not required in other types of research. Member checking of the analysis, however, is not practical. If the participant does not agree with the analysis, this must place the researcher in an awkward position. Recall that the analysis is usually a synthesis of all of the interviews, and the text will have been abstracted; therefore, it is highly unlikely that a participant would recognize their own story in the combined text. Of greater concern, the text may have been analyzed interpretively. Thus, if the participant does not think the analysis is right, what does the investigator do? Change the text to whatever the participant thinks it should be? Discard that part of the analysis? What does the researcher do if the participant suggests changes? The researcher's background in theory and research methods must outrank the participant as a judge of the analysis. Therefore, member checking as a strategy is not recommended.

External Audits

External audits are best used as a system of *auditing*, and should, therefore, be conducted when the researcher's findings are suspect. Perhaps a granting agency suspects that the results are not as balanced as they should be, that the results are "too good" to be true, or that some type of undesirable investigator bias is present. External auditors would, therefore, request to see the researcher's data, explore the processes of conceptualization, and reexamine the investigator's conclusions.

The findings of a qualitative researcher are rarely challenged. One high-profile case was Margaret Mead's (1928) work in Samoa, which was challenged by Freeman in 1983. This led to one of the most controversial debates in the literature (see, for instance, Ember, 1985; Feinberg, 1988; Grant, 1995; Hill, 1992; Shankman, 2009)—a debate centered on validity of ethnographic fieldwork.

Triangulation

Triangulation has a number of levels—investigator, data, theory, or methods. For establishing validity, it usually refers to the use of two or more sets of data or methods to answer one question. This process increases scope or depth of the study, because different sets of data or different qualitative methods may each elicit different data, different participants, or perspective, and it may beg the question, "Which data set or methods is the correct one?"

However, with the difficulty of collecting qualitative data or conducting a qualitative method, a larger question arises: "Why conduct the same study twice?" "Does one not have adequate trust in the first method, such as to go to the bother of duplicating the study using a second method?" Qualitative studies take much time and effort, so going to the additional work of duplicating the study in a way that gives little gain (i.e., determining validity) appears pointless. Finally, researchers do not usually publish both studies in the same journal, which would allow the reader to judge whether both studies are the same. If, however, the second study was conducted to gain information that the first method did not provide, then a second study—a multiple-method design—may be worthwhile.

An interesting attempt at using the same data with different methods was conducted using the same data with five different qualitative methods (Wertz et al., 2011). The different lens of each method (phenomenology, grounded theory, discourse analysis, narrative inquiry, and intuitive inquiry), and a chapter comparing the results of each analysis, provides an important example of the value of a multiple-method project and the different understandings gained from each perspective and method. This variation endorses the significance of the research basing their study on an overt theoretical foundation or context. It provides the reader with some theoretical context with which to judge the validity of the results.

Summary

The strategies for attaining validity are summarized in Table 1. Of importance, member checking is not recommended, and

Table 1. Summary of Recommendations for Establishing Validity in Qualitative Inquiry.

Strategy	Validity	Comments and Caveats
Prolonged engagement	Necessary for observational research to reduce "observer effects"	Not necessary for interview research
Observation	Prolonged observation reduces observer effect	Spot observation sampling will reduce "observer effects"
Thick description	Unstructured interviews An adequate and appropriate sample necessary	Attend to indices of saturation
	For unstructured interviews: A large sample is necessary to obtain adequacy	Description enhanced in semi-structured interview research if probes are used
Triangulation	Methodological triangulation may enhance validity as multiple-method research	It is used to expand understanding
Development of a coding system and inter-rater reliability	Use only for semi-structured research	Will invalidate research that uses unstructured open-ended and guided interview
Researcher bias	1. Unconscious bias: may be evident in the question and design 2. May be used beneficially to maximize phenomena	1. May warp findings 2. May validate findings
Negative case analysis	Necessary for research using unstructured interviews	With semi-structured interviews, attend to questions that have not been answered
Peer review debriefing	May be used to assist the researcher conceptualize	If researching using unstructured interviews
Member checking	Not used	
External audits	Not routinely used	Problematic—too late to fix identified concerns

all of the other strategies have caveats, limiting their application.

Exploring Strategies to Ensure Reliability

In qualitative inquiry the major strategies for determining reliability occur primarily during coding. Inter-rater reliability is the comparison of the results of a first coder and a second coder. Reliable coding is demonstrated by duplication, which is ensured by explicit coding decisions, communicated by clear definitions in a codebook, so that a trained second coder may make the same coding decisions as the first coder. Reliable studies are stable and not subject to random variation. Therefore, the major strategies to ensure *reliability* are the development of a coding system and inter-coder agreement. Other strategies for reliability include member checks and peer review debriefing, triangulation, and external audits.

Development of a Coding System

The development of a coding system for use with semi-structured interviews is the listing of all possible responses (or codes) to all items (see, McIntosh & Morse, in press). When preparing semi-structured interviews, questions are pre-selected within a delimited domain, and the format of the interviews contains the stems of these fixed questions.

This set of questions is asked of all participants, in the same order. Participants are free to answer each question as they wish, but answers are restricted by the question's stem, and become patterned relatively quickly, although the investigator may ask follow-up probes to obtain additional information. These follow-up questions may or may not be included in the coding scheme.

The coding of these responses for semi-structured interviews is limited to rather short statements about a very restricted topic. The structure and order of the interviews and the questions do not change. The responses to these questions are then summarized into a clearly delimited set of responses that are numbered and well described. Thus, a second coder may recognize the descriptions and code using the same numbers (codes) for the same item, and make the *same coding decisions* as the first coder, using any one piece of text.

This standardization enables the responses to be systematically coded—and also numerically coded—item-by-item. Inter-rater reliability is the obtaining of a second coder to recode data—to see whether that person obtains the same results as the first coder. The second coder usually randomly selects sections of data to recode; rarely is the entire data set double coded (Morse, 1997). Inter-coder reliability statistics may be used to numerically determine the accuracy of such a coding system, thereby demonstrating the stability of the coding and the results.

With *un*structured interviews, however, this system cannot be used, and it actually invalidates the research. Unstructured interviews are narrative interviews, in which, in response to a general question, the participant "tells his story" without interruption. Not only is the interview unstructured, but also research design in ethnography and grounded theory demand that the researcher is learning about the phenomenon as the study progresses. Subsequently, unstructured interviews change as the researcher becomes more knowledgeable about the topic, so that one interview does not cover exactly the same material as the next interview. Later interviews may be more detailed about some particular aspect; they may be used to confirm data from other interviews, or even to test conjectures. Therefore, coding decisions are not made from patterned replication (as with semi-structured interviews), but interpretively, within the context of all of the previous interviews. Pieces of the interview are interdependent, one portion enlightening the researcher's understanding about other portions. The first coder codes according to what data signify. Later, coding decisions are made in the context of the knowledge or information gained from all of the interviews. *This is the essence of interpretive coding.*

Thus, small pieces of data may have important significance to a researcher who is coding and intimately familiar with the total data set, but the significance of such cues and clues will be missed by the second coder, because the second coder has not listened intently to the entire interview, or set of interviews. The second coder is making the coding decision entirely on the piece of text before him (rather than the interview or data set as a whole), and does not, and cannot, code as interpretively as the first coder. Keeping the coding reliable occurs at the expense of interpretive insight. Using inter-rater reliability with unstructured interviews keeps the coding system descriptive and inhibits interpretive coding. The use of a second coder will keep the analysis superficial, trivial, obvious, insignificant, uninteresting, and trite. In other words, while the use of multiple coders enhances *demonstrated reliability* for semi-structured research, it *invalidates* research that uses unstructured interviews.

What is "demonstrated reliability"? Demonstrated reliability occurs when a study uses thick, rich and saturated data, data that are verified in each step of data collection, and has data that represent the phenomenon (either concrete or experiential and abstract) to such a degree that others may instantly recognize it (Morse, in press-a). Demographic tables in qualitative studies should list the reasons for selecting participants in the study, rather than reporting irrelevant demographic characteristics (Morse, 2008b), thereby providing information about the sampling rationale.

Member Checks[4]

If your analysis was detailed, rich, and descriptive, any reliability checks and balances will have been conducted in the course of verification during concurrent data collection and analysis. Member checking is more than returning an interview to participants to check that the transcription is correct. Member checking may be conducted during the process of data collection to check data between participants: "Other people tell me *[thus and so]*. Is this how it is for you?" Notice that this member checking is not done with the original participants, but with others. Such replication determines normative patterns of behavior, hence achieves reliability.

Does the researcher understand/interpret the participant correctly? In the text, as a reliability issue, if you do not understand what is going on, your analysis is *unstable, and the results cannot be repeated*, therefore, it is impossible to get reliable results.

Thick Description

When data are plentiful, data naturally overlap. Of course, examples are not exactly the same, but key issues bear resemblances. This overlap enables the researcher to see replication; hence, thick description contributes to internal reliability.

Table 2. Summary of Recommendations for Establishing Reliability in Qualitative Inquiry.

Strategy	Reliability	Comments and Caveats
Development of a coding system and inter-rater reliability	Only for semi-structured interview research	Coding system and codebook are essential
Member checks	Does the researcher understand/interpret the participant correctly?	In text: As a reliability issue, if you do not understand what is going on, your analysis is *unstable, and cannot be repeated* to get the same results.
Thick description	Provides opportunity for seeing replication/ duplication	Interviews overlap, and therefore, verify the data set internally.
Peer review debriefing	Not usually a reliability issue, except for team research	
External audits	Do not ensure reliability Not routinely used	Problematic—too late to fix identified concerns

Triangulation

Triangulation as a strategy for reliability is very closely associated with using triangulation as a strategy for validity. For processes of data triangulation to provide reliable results, analysis of data must provide similar results. This may be done by splitting the data into two sets and having two investigators analyze each set (of course, making certain that the data in each pile is adequate), or asking two investigators to independently analyze the same data—which, of course, may be cumbersome.

If the researcher uses two *methods*, the researcher again examines the results of the two methods, asking, "Do both methods give the same results?" If the same results are obtained with each analysis, theoretically, the results are reliable. But, of course, you recognize that two different methods use two different perspectives and will not give the same results. Furthermore, if there are two researchers, their skill in interpreting the data is a large factor in the nature and quality of the findings, even when both are using the same research question, literature, and data.[5] Therefore, obtaining the same results independently, simply to demonstrate reliability, is rather a lost cause and not recommended.

External Audits

External audits may reveal internal linkages as the theory develops, but it is not a routine strategy, and it is of limited use as a tool to achieve validity, not reliability.

Summary

Reliability is a well-developed concept for semi-structured interviews that have a relatively rigid structure and guidelines for use. Determining reliability for research using unstructured interviews is managed internally as data accrues, and the similarity between data from different participants is observed. Other strategies, summarized in Table 2, are not practical or well suited to determine reliability.

Recommendations

What does this all mean for determining the rigor of qualitative inquiry? First, it means that rigor, comprising both validity and reliability, is achieved primarily by researchers in the process of data collection and analysis. Most of the strategies are researcher centered (see Table 3): prolonged engagement, thick description, development of the coding system and inter-rater reliability, the correction of researcher bias, negative case analysis, and peer review/ briefing. Member checking is conducted with participants during the analytic process, and external audits are conducted by "auditors" following the completion of the project, and are not intended for routine use by investigators.

From this critique, it is clear that many of the strategies do not achieve their intended goal of contributing to rigorous research, and must be used judiciously. Member checking is of little value to determine validity and does not achieve reliability. Some strategies may be used only for unstructured interviews (i.e., thick description, negative case analysis, researcher bias, and perhaps peer review), and some only used for semi-structured research (i.e., use of a coding system and inter-rater reliability). Some strategies may only be used with particular methods, such as prolonged engagement with observational research, or methodological triangulation with mixed-method designs. The indiscriminate use of strategies with any type of qualitative research is clearly harmful to the goal of conducting rigorous research.

Table 3. Summary of Recommendations for Strategies for Establishing Rigor in Qualitative Inquiry.

Strategy	Validity	Reliability
Prolonged engagement	Yes: for research using observation	No
Thick description	Yes: for research using unstructured interviews	More opportunity for replication of data
Triangulation	Yes: for mixed-method research	No
Development of a coding system and inter-rater reliability	Yes: only for semi-structured interview research	Yes: essential for semi-structure interview research and multiple coders
Researcher bias	May be evident in research question and design (groups not equivalent etc.) Data will correct themselves if researcher is responsive to the principles of induction	Not a reliability concern
Negative case analysis	Yes: for research using unstructured interviews With semi-structured interviews, attend to missing responses	Not used as a reliability measure
Peer review/debriefing	Yes: may assist with conceptualization	Not used
Member checking	Not used	Not used
External audits	Not routinely used	Not routinely used

An additional problem occurs with external evaluation of rigor following the completion of the research (Sparkes & Smith, 2009). The "audit" is not practical for use by journal or grant reviewers or readers, perhaps interested in implementing or teaching the research. The Guba and Lincoln criteria are only useful if the reviewers note that the strategies have been done during the course of the research and reported in the article—they are not evident in the quality of the completed research. As a result, new evaluation criteria are emerging that appears to be more useful for post-completion rating, such as the criteria proposed by Tracy (2010); Cohen and Crabtree (2008). These criteria often are less specific than the Guba and Lincoln criteria (for instance, focusing on coherence, contribution, and significance of the research rather than specific strategies). The problem with such criteria is that they largely remain subjective, and do not, for instance, assist with the "quality" questions that are demanded by those who, for instance, are conducting meta synthesis.

There is still much work to do in this area of rigor. We do not know the extent to which these strategies are used incorrectly, or non-specifically, and the price to qualitative inquiry, in research time, research costs, and in the diminishing quality of our research. However, rather than focusing on such problems, I recommend that we develop, refine, and test analytic processes and strategies that fit qualitative inquiry while also remaining consistent with concepts used by the larger social science community. Only then will we be able to describe our methods in a way that other social sciences will comprehend and respect our research.

Author's Note

This article was presented as a keynote address at the 4th Global Congress for Qualitative Health Research as "Reformulating Reliability and Validity in Qualitative Inquiry," Mérida, Mexico, March 18, 2015.

Declaration of Conflicting Interests

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

Funding

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

Notes

1. In qualitative inquiry, because of sample size and the comparatively cumbersome modes of data collection and analysis, the practice of randomization would decrease

validity, as the chances of gaining understanding, so essential in theoretical sampling, would be lost. When quantitative researchers consider theoretical sampling as a source of bias, they are removing the practice from the checks and balances of the process of data collection and analysis as a whole (Morse, 2006).

2. This is another issue that differentiates qualitative inquiry from quantitative, where in qualitative inquiry validity issues are primarily applied to measurement (M. A. Carey, personal communication, May 4, 2015).
3. This point, that some of these strategies for ensuring trustworthiness may be used for some strategies or methods, but not for others, has not been noted by other researchers. For instance, Creswell (2012) states that multiple coders are "needed for highly interpretive texts" (p. 253). This is incorrect.
4. Peer review debriefing described by Guba refers only to validity. It is not clear why Creswell (2012) included it as a strategy for determining reliability.
5. Wertz et al. (2011) attempted to demonstrate the differences of five qualitative methods, using the same qualitative data, and then comparing their findings with the participant's analysis. Unfortunately, the demonstration was not, as you can imagine, a "success." "We six authors, leave you, as reader, to draw your own conclusions . . ." (p. 367). This is also another example of the "Blind men and the elephant" parable, and different conclusions being reached by examining different parts of the whole.

References

Agar, M. (1996). *The professional stranger*. San Diego, CA: Academic Press.

Cohen, D. J., & Crabtree, B. F. (2008). Evaluative criteria for qualitative research in health care: Controversies and recommendations. *The Annals of Family Medicine*, *6*, 331–339.

Corbin, J., & Morse, J. M. (2003). The unstructured interactive interview: Issues of reciprocity and risks. *Qualitative Inquiry*, *9*, 335–354.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.

Ember, M. (1985). Evidence and science in ethnography: Reflections on the freeman-mead controversy. *American Anthropologist*, *87*, 906–910.

Feinberg, R. (1988). Margaret Mead and Samoa: Coming of age in fact and fiction. *American Anthropologist*, *90*, 656–663.

Freeman, D. (1983). *Margaret Mead and Samoa: The making and unmaking of an anthropological myth*. Cambridge, MA: Harvard University.

Grant, N. J. (1995). From Margaret Mead's field notes: What counted as "sex" in Samoa? *American Anthropologist*, *97*, 678–682.

Guba, E. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal*, *29*, 75–92.

Guba, E., & Lincoln, Y. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Guba, E., & Lincoln, Y. (1989). *Fourth generation evaluation*. Newbury Park, CA:Sage.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, *18*, 59–82.

Hill, J. D. (1992). Contested pasts and the practice of anthropology. *American Anthropologist*, *94*, 809–815.

Krefting, L. (1991). Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy*, *45*, 214–222.

Maxwell, J. A., & Chmiel, M. (2014). Generalization in and form qualitative analysis. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 541–553). Thousand Oaks, CA: Sage.

McIntosh, M., & Morse, J. M. (in press). Situating and constructing the semi-structured interview. *Global Qualitative Nursing Research*.

Mead, M. (1928). *Coming of age in Samoa: A psychological study of primitive youth for Western civilization*. New York: William Morrow.

Meadows, L. M., & Morse, J. M. (2001). Constructing evidence within the qualitative project. In J. M. Morse, J. M. Swanson, & A. Kuzel (Eds.), *The nature of qualitative evidence* (pp. 187–202). Thousand Oaks, CA: Sage.

Miller, P. (2008a). Reliability. In L. Given (Ed.), *Encyclopedia of qualitative methods* (p. 745). Thousand Oaks, CA: Sage.

Miller, P. (2008b). Validity. In L. Given (Ed.), *Encyclopedia of qualitative methods* (p. 909). Thousand Oaks, CA: Sage.

Morse, J. M. (1997). Perfectly healthy, but dead: The myth of inter-rater reliability [Editorial]. *Qualitative Health Research*, *7*, 445–447.

Morse, J. M. (1998). Validity by committee [Editorial]. *Qualitative Health Research*, *8*, 443–445.

Morse, J. M. (2001b). Using shadowed data [Editorial]. *Qualitative Health Research*, *11*, 291–292.

Morse, J. M., & Mitcham, C. (2002). Exploring qualitatively-derived concepts: Inductive-deductive pitfalls. *International Journal of Qualitative Methods*, *1*(4), Article 9. Retrieved from http://www.ualberta.ca/~ijqm.

Morse, J. M. (2006). Biased reflections: Principles of sampling and analysis in qualitative enquiry. In J. Popay (Ed.), *Moving beyond effectiveness in evidence synthesis: Methodological issues in the synthesis of diverse sources of evidence* (pp. 53–60).

Morse, J. M. (2008a). Does informed consent interfere with induction? [Editorial]. *Qualitative Health Research*, *18*, 439–440. doi:10.1177/1049732307313614

Morse, J. M. (2008b). "What's your favorite color?" Irrelevant demographic detail in qualitative articles [Editorial]. *Qualitative Health Research*, *18*, 299–300. doi:10.1177/1049732307310995

Morse, J. M. (2010). "Cherry Picking": Writing from thin data [Editorial]. *Qualitative Health Research*, *20*, 3. doi:10.007/104973230935428

Morse, J. M. (in press-a). Building validity in qualitative research. *Journal of Qualitative Research*.

Morse, J. M. (in press-b). Reframing rigor in qualitative inquiry. In N. Denzin & Y. Lincoln (Eds.), *SAGE handbook of qualitative inquiry* (5th ed.). Thousand Oaks, CA: Sage.

Morse, J. M. (in press-c). Towards understanding comfort & comforting. In *Analyzing and constructing the conceptual and theoretical foundations of nursing*. Philadelphia: F. A. Davis.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods*, *1*(2), 13–22.

Nussbaum, E. (1998). Return of the natives: What happens when an anthropologist's manuscript is edited by his subjects? *Lingua Franca*, *8*(1), 53–56.

O'Neill, T. (1995). Implementation frailties of Guba and Lincoln's fourth generation evaluation theory. *Studies in Educational Evaluation*, *21*, 5–21.

Polit, D. F., & Beck, C. T. (2012). *Nursing research: Generating and assessing evidence for nursing practice* (9th ed.). Philadelphia: Lippincott, Williams & Wilkins.

Popper, K. R. (1965). *Conjectures and refutations: The growth of scientific knowledge*. New York: Harper Torchbooks. (Original work published 1963)

Shankman, P. (2009). *The trashing of Margaret Mead: Anatomy of an anthropological controversy*. Madison: The University of Wisconsin.

Sparkes, A. C., & Smith, B. (2009). Judging the quality of qualitative inquiry: Criteriology and relativism in action. *Psychology of Sport and Exercise*, *10*, 491–497.

Spiers, J. (in press). The pink elephant paradox (or, avoiding the misattribution of data). In J. Morse (Ed.), *Analyzing and constructing the conceptual and theoretical foundations of nursing*. Philadelphia: F. A. Davis.

Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry*, *16*, 837–851.

Tuckett, A. (2005). Part II: Rigour in qualitative research: Complexities and solutions. *Nurse Researcher*, *13*(1), 29–42.

Wertz, F. J., Charmaz, K., McMullen, M., Josselson, R., Anderson, R., & McSpadden, E. (2011). *Five ways of doing qualitative analysis phenomenological psychology, grounded theory, discourse analysis, narrative research, and intuitive inquiry*. New York: Guilford Press.

Author Biographies

Janice M. Morse, PhD, is a professor and Barnes Presidential Chair, University of Utah, and professor emeritus, University of Alberta, Canada.