Theory to Practice

What Can You Tell From An N of 1?: Issues of Validity and Reliability in Qualitative Research

Sharan B. Merriam

Abstract

This article deals with the issues of validity and reliability in qualitative research in education. Philosophical assumptions underlying the concepts of internal validity, reliability, and external validity or generalizability are discussed. In addition, strategies congruent with a qualitative research perspective for ensuring for rigor and trustworthiness of findings in qualitative research are presented.

At conference presentations, in reviews of journal articles, at thesis defenses, the trustworthiness of qualitative research continues to be challenged, and rightly so. Rigor is needed in all kinds of research to insure that findings are to be trusted and believed. In applied fields like education, social work, counseling, and administration, the question of the trustworthiness of research findings looms large; after all, much research is designed to understand and improve practice. We want to feel confident incorporating research findings into our practice, for what we do affects the lives of real people.

Questions most commonly posed to qualitative researchers reflect concerns with the validity and reliability of the research findings—questions such as the one in the title of this article, and others such as: “How can you generalize from a small, non-random sample?” “If somebody else did this study, would they get the same results?” “How do you know the researcher isn’t biased and just finding what he or she expects to find?” and “If the researcher is the primary instrument for data

Sharan Merriam is Professor of Adult Education at the University of Georgia.
collection and analysis, how can we be sure the researcher is a valid and reliable instrument? These questions reflect legitimate concerns about the rigor of qualitative research; they also reflect philosophical assumptions underlying a quantitative or positivist worldview and are thus inappropriate for assessing the rigor of a qualitative study. The purpose of this paper is twofold: (1) to examine conceptions of validity and reliability from a qualitative or interpretive worldview and (2) to present strategies for ensuring for validity and reliability that are consonant with assumptions underlying the qualitative paradigm.

Purposes of Qualitative Research

In assessing the trustworthiness of qualitative research, it is important to back up and ask what kinds of questions or problems qualitative research is designed to address. Qualitative research is ideal for the following: clarifying and understanding phenomena and situations when operative variables cannot be identified ahead of time; finding creative or fresh approaches to looking at over-familiar problems; understanding how participants perceive their roles or tasks in an organization; determining the history of a situation; and building theory, hypotheses, or generalizations. The question of trustworthiness becomes how well a particular study does what it is designed to do.

Notions of validity and reliability must be addressed from the perspective of the paradigm out of which the study has been conducted. That is, if I am trying to build hypotheses rather than test them, if I am trying to understand a phenomenon rather than "test it," if I am interested in participants' perspectives rather than my own, different questions will need to be asked about the conduct of the study. Qualitative researchers have approached rigor from one of two angles. Some lay out the standard, positivist threats to validity and reliability made famous by Campell and Stanley (1966) and Cook and Campbell (1979) and demonstrate how qualitative research addresses these threats. History, maturation, observer effects, selection and regression, mortality, spurious conclusions, and so on, can be assessed from a qualitative research perspective as demonstrated by Guba and Lincoln (1981) and Goette and LeCompte (1984). More commonly, writers make the case that qualitative research is based on different assumptions regarding reality, thus demanding different conceptualizations of validity and reliability. Some have proposed using a different nomenclature. Agar (1986), for example, talks about credibility, accuracy of representation, and


The position expressed in this paper is that notions of validity and reliability need to be grounded in the worldview of qualitative research. Further, there are strategies that can be employed to ensure for trustworthiness that are highly compatible with this worldview. The discussion that follows is presented in terms of the three major aspects of rigor—internal validity, reliability, and external validity (generalizability).

Internal Validity

Internal validity asks the question: How congruent are the findings with reality? In qualitative research the question is often more precisely stated: Are we observing or measuring what we think we are observing or measuring? Key to understanding internal validity is the notion of reality. Is reality fixed and stable as the positivists believe, or constrained and interpreted as qualitative researchers believe? These two views of reality are eloquently contrasted in Steinbecks's (1941) log of his scientific journey to the Sea of Cortez:

We knew that what we would see and record and construct would be warped, as all knowledge patterns are warped, first, by the collective pressure and strain of our time and race, second by the thrust of our individual personalities. But knowing this, we might not fall into too many holes—we might maintain some balance between our warp and the separate things, the external reality. The oneness of these two might take its contribution from both. For example, the Mexican sierra has "XVII-15-IX" spines in the dorsal fin. These can easily be counted. But if the sierra strikes hard on the line so that our hands are burned, if the fish sounds and nearly escapes and finally comes in over the rail, his colors pulsing and his tail beating the air, a whole new relational externality has come into being—an entity which is more than the sum of the fish plus the fisherman. The only way to count the spines of the sierra unaffected by this second relational reality is to sit in a laboratory, open an evil smelling jar, remove a stiff, colorless fish from formalin solution, count the spines, and write the truth "D. XVII-15-IX." There you have recorded a reality which cannot be ascribed—probably the least important reality concerning either the fish or yourself. It is good to know what you are doing. The man with his pickled fish has set down one truth and has recorded in his experience
Reliability is concerned with the question of the extent to which two or more findings with the same construct would be expected to be stable over time. This is another way of phrasing the idea of replicability. Typically, investigators distribute the same criteria to different cases in different settings and at different times. Typically, the results of such studies are reliable if the findings are stable over time, even when different people and different settings are involved.

Suppose, for example, that you were to study the relationship between stress and job performance. You might find that employees who have high levels of stress also have low job performance. If you were to conduct the same study at a different time or in a different setting, you might find that the same relationship holds. This would indicate that your findings are reliable. On the other hand, if you were to conduct the same study at a different time or in a different setting and find that the relationship does not hold, you might conclude that your findings are not reliable. This is another way of saying that the results of your study are not consistent over time. This is an important consideration in any scientific research, and it is one that is often overlooked.

Reliability can be measured in different ways, depending on the type of research being conducted. For example, reliability can be measured by the correlation between two or more sets of data. This is known as the coefficient of reproducibility. Another way of measuring reliability is to compare the results of different studies that have been conducted in similar settings. This is known as the coefficient of reproducibility. These measures can be used to determine the extent to which the results of a study are consistent over time.
Scriven (1972) makes the point that a lot of people experiencing the same thing does not necessarily mean that their accounts are more reliable than that of a single individual. Five hundred people reporting that they had seen a magician cut a person in half, for example, would not be as reliable as a report as that of the lone standbag who had witnessed the event from behind the curtain.

Qualitative researchers are not seeking to establish "laws" in which reliability of observation and measurement are essential. Rather, qualitatively researchers seek to understand the world from the perspectives of those in.

Since there are many perspectives, and many possible interpretations, "there is no benchmark by which one can take repeated measures and establish reliability in a traditional sense" (Merriam, 1988, p. 170). Clearly, replication of a qualitative investigation will not yield the same results (as it might in quantitative research). Rather, both sets of results stand as two interpretations of the phenomenon.

Instead of reliability, one can strive for what Lincoln and Guba (1985, p. 288) call "dependability," or "consistency." The real question for qualitative researchers, they suggest, is not whether the results of one study are the same as the results of a second or third study, but whether the results of a study are consistent with the data collected. As with internal validity, there are strategies one can use to ensure for greater consistency. Three such strategies are listed below:

1. Triangulation—The use of multiple methods of data collection, in particular, as well as other forms of triangulation, can lead to dependability or consistency (as well as internal validity).

2. Peer examination.—Again, this strategy provides a check that the investigator is plausibly interpreting the data; that is, someone else can be asked whether the emerging results appear to be consistent with the data collected.

3. Audit trail.—This strategy, suggested by Guba and Lincoln (1981), operates on the same premise as when an auditor verifies the accounts of a business. "In order for an audit to take place, the investigator must describe in detail how data were collected, how categories were derived, and how decisions were made throughout the inquiry" (Merriam, 1988, p. 172). Goetz and LeCompte (1984) suggest that the audit trail should be so detailed "that other researchers can use the original report as an operating manual by which to replicate the study" (p. 216).

Reliability, then, cannot be thought of in qualitative research in the same way as it is in positivist research. The logic of reliability in quantitative research is based on philosophical assumptions and a worldview.

Merriam

different from that of qualitative research. What one strives for is consistency and dependability, a sort of internal reliability in which the findings of an investigation reflect, to the best of the researcher's ability, the data collected.

External Validity

The extent to which the findings of a study can be applied to other situations refers to the question of external validity, or generalizability. Indeed, this question seems to haunt qualitative research more than any other, probably because most people think of generalizability in the statistical sense of extrapolating from a sample to a population. Since qualitative researchers rarely select a random sample (which would then allow them to generalize to the population from which the sample was selected), it is thus concluded that one cannot generalize in qualitative research. However, even generalizability in the statistical sense has its limitations. First, although random sampling may have been used, even generalizations are made within certain levels of confidence. Outlying scores are ignored (they become part of the variance). Second, when dealing with human beings, it is difficult, if not impossible, to apply statistically-based generalizations to individual persons. For example, a study might find that 70% of the surgeons in the Northeast voted Republican in the last election. But any individual surgeon, say a Dr. Jones, will have voted no/70% Republican, but either 100% Republican or 100% Democrat. Third, several writers have pointed out that very little social science research is based on true random sampling of a population; most populations from which samples are selected are confined to small geographic areas, homogeneous groups, or highly defined subgroups (middle-school principals in a single state, for example).

While some qualitative researchers view generalizability as a limitation of the method, or just not appropriate for the social sciences, most prefer to think of generalizability as something different than going from a sample to a population. The goal of qualitative research, after all, is to understand the particular in depth, rather than finding out what is generally true of humanity. There are at least three alternative conceptions of generalizability that are congruent with the philosophical assumptions underlying qualitative research. Creswell (1975) thinks that empirical generalizations are true, and a goal for social science research; rather, we should think in terms of working hypotheses. He writes: "Instead of making generalization the ruling consideration in our research, I suggest
Theory to Practice

that we reverse our priorities. An observer collecting data in one particular situation (as is common in experimental research) is quite likely to regard the situation as different and to interpret the data in different ways. In this respect, the results of such experiments have limited validity. Generalization comes late. When we give equal weight to local conditions, any generalization is a working hypothesis, premature, unsatisfactory, even specious (p. 124-125).

Working hypotheses are fragile constructs. They can be used to guide practice (Panitz, 1997).

A somewhat different concept, called concrete universals, has been proposed by Edward E. Ericsson (1986). In attending to the particular, universals can be discovered.

Concrete universals are based on the notion that particular situations are not studied in isolation from other situations. Rather, the ideas are drawn from a variety of situations that share the same general characteristics (p. 182). For example, a teacher who observes a student behaving in a certain way in one setting can generalize that the student is likely to behave similarly in other settings. This allows for a more flexible and adaptable approach to teaching, as teachers can recognize patterns of behavior across different contexts and adjust their strategies accordingly.

A third way of viewing external validity is something becoming known as teacher or user generalizability. In this view, the extent to which findings from an investigation can be applied to other situations is determined by the people in those situations. It is not up to the researchers to determine the extent to which their findings can be applied to other settings; it is up to the consumers of the research. Wilson (1970) suggests the notion of a continuum of usefulness beginning on one end representing the setting where the information was gathered and stretching in diminutive settings. (p. 434)

Whether one thinks of generalizability in terms of working hypotheses, concrete universals, or user generalizability, there are strategies one can employ to strengthen this aspect of rigor in qualitative research. Four such strategies are discussed below:

1. Thick descriptions. This involves providing enough information about the phenomenon under study so that readers will be able to determine how closely their situations match the research situations, and whether their findings can be transferred.
2. Multiple case designs. The use of several sites, cases, situations, especially those representing some variation (Glaser & Strauss, 1967), will allow the results to be applied to a greater range of other similar situations.

References

3. Mixed control—This strategy involves describing how typical the group or event is, as compared with the majority of others in the same class. In Wolcott’s (1973) case study of a school principal, for example, he tells the reader how representative his subject is compared to the typical school principal.

4. Sampling within—A phenomenon being studied may have numerous component parts (teachers, administrators, students in a school system, for example), each of which could be randomly sampled for inclusion in the study. This would allow one to “generalize” to the larger group within the unit of study.

In summary, external validity or generalizability seems to be most problematic for those too well acquainted with qualitative research. To consider generalizability a limitation of this kind of research is to be thinking in terms of statistical generalization based on the quantitative paradigm. By viewing external validity from the perspective of the assumptions underlying qualitative research, several reformulations of generalizability are possible, such as working hypotheses, concrete universals, and teacher or user generalizability.

What Can You Tell From an N of 1?

Viewed from a qualitative perspective, quite a bit can be learned from an N of 1. The usefulness of the findings of a study with a small N and no random sampling is dependent upon the internal validity, reliability, and external validity of the study. As was discussed in this article, there are ways to view each of these concerns that are congruent with the underlying assumptions and world-view of qualitative research. Likewise, there are strategies that investigators can employ that will ensure for the validity and reliability of the study. Rigor is a valid concern in qualitative research as in any other kind of research. Qualitative researchers employ different means of “persuading” the reader that a study is trustworthy. This is what Frampton (1987) calls the "heterotic" of this research. While the qualitative study must convince the reader that procedures have been followed faithfully because very little concrete description of what anyone did is provided, qualitative research pertains through its "classical ultimate of concrete description of detail, portrayal of process in an active mode, and attention to the perspectives of those studied" (pp. 19, 20).
References