International Journal of Qualitative Methods

INTERNATIONAL INSTITUTE FOR QUALITATIVE METHODOLOGY (IIQM)
UNIVERSITY OF ALBERTA

Article

Reconsidering Evaluation Criteria for Scientific Adequacy in Health Care Research: An Integrative Framework of Quantitative and Qualitative Criteria

Hiroaki Miyata, PhD
Department of Healthcare Quality Assessment
Graduate School of Medicine
The University of Tokyo, Japan

Ichiro Kai, MD, MPH
Department of Social Gerontology
School of Public Health
The University of Tokyo, Japan

Abstract

It is important to reconsider evaluation criteria regarding scientific adequacy in health care research. In this article the authors review the four pairs of quantitative/qualitative paradigms. They discuss the use of evaluation criteria based on a pragmatic perspective after examining the epistemological issues behind the criteria. Validity/credibility is concerned with research framework, whereas reliability/dependability refers to the range of stability in observations, objectivity/confirmability reflects influences between observers and subjects, and generalizability/transferability has epistemological differences in the way findings are applied. Qualitative studies should not always choose qualitative paradigms, and vice versa. If stability can be assumed to some extent in a qualitative study, it is better to use a quantitative paradigm. Regardless of whether it is quantitative or qualitative research, it is important to recognize the four epistemological axes.

Keywords: validity, reliability, generalizability, quantitative research, qualitative research, evaluation

Historical relationship between quantitative and qualitative research

Newton's era and the Renaissance established two permanently important worldviews from which modern traditions and methods of inquiry have proceeded. These two worldviews and the position of scientists within them roughly characterize quantitative and qualitative research. The quantitative world view posits an objective reality characterized by stable, predictable, and commensurable phenomena that operate through the laws of cause and effect. The qualitative worldview emphasizes a subjective reality that is characterized by complexity, apparently infinite variation, and incommensurability without perturbation. The laws of cause and effect operate, but always within a unique set of circumstances determined by multiple factors (Thomas, 1996).

Historically it might have been difficult for many health care practitioners and researchers, the majority with educational backgrounds strictly in the natural sciences or biology, to take a qualitative approach because qualitative methods are rooted in social science (Pope & Mays, 2001). The wide acceptance of qualitative research might also have been slowed because of the influence of views such as those of Kuhn (1970), and Bernstein and Freeman (1975), for example, who believed quantitative prediction to be preferable to qualitative prediction. They considered research that had both quantitative and qualitative data to be less valuable in terms of method than research that employed quantitative data only.

The scientific community was often seen as divided into two groups based on the use of quantitative versus qualitative methods, and the schism between these seemingly conflicting traditions seemed profound and unlikely to be bridged. In recent years, however, in the fields of both social sciences and health care, researchers have come to believe that it is no use to rigidly separate between quantitative and qualitative research (Abell, 1990; Barbour, 1999; Hammersley, 1992; Mechanic, 1989; Pearlin, 1992).

The realization that quantitative and qualitative methods are not incompatible was an important one. Although the two approaches are useful on their own, they are also powerful when used in conjunction with each other. For example, in the form of a questionnaire, qualitative research can be used to narrow a research focus in preparation for use of quantitative methods. Qualitative research is also useful for deeper interpretations of quantitative findings. On the other hand, large variances in quantitative analysis of response data can be explained by focused interviews. Mixed-method research (Johnson & Onwuegbuzie, 2004; O'Byrne, 2007), in which the researcher uses the qualitative research paradigm for one phase of a research study and the quantitative research paradigm for another phase of the study, is also important in health care. Systematic reviews of multimethod studies have demonstrated that significant results would not be possible without equal evaluation of both quantitative and qualitative data (Bravata, McDonald, Shojania, Sundaram, & Owens, 2005; Mulrow, Langhorne, & Grimshaw, 1997). These findings intimate the need to create new integrative criteria by which to evaluate the scientific adequacy of both quantitative and qualitative analysis.

Development of the debate of a an integrative framework for evaluation

There are three main perspectives—method, use, and value—regarding evaluation criteria (Alkin, 2004). As scientific adequacy is one of the most important criteria in method, it was well organized in the social experiment field. Although many researchers recognize the importance of scientific adequacy in qualitative research, it was not well organized in qualitative research field until now. Recently some researchers have developed criteria including not only method but also use or value (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Messick, 1995). In this study, however, we focus on method and

examine an integrative framework of quantitative and qualitative criteria. There are many words and concepts (Koch, 1994; Watson & Girard, 2004) regarding method in qualitative research field (e.g., rigor, integrity); we use *scientific adequacy* in this study.

Validity, reliability, and generalizability are widely used as criteria for the evaluation of quantitative analysis. However, many qualitative researchers, who do not assume an objective reality or a confirmatory perception, tend to question this holy trinity (Klave, 1995). Even now, more than two decades since it began, the interdisciplinary discussion regarding the roles of validity, reliability, and generalizability in the qualitative paradigm continues (Mays & Pope, 2001). Debate concerning the relationship between quantitative and qualitative paradigms is often muddled and confused, and the clutter of terms and arguments has resulted in the concepts' becoming obscure and unrecognizable (Morse, Barrett, Mayan, Olson, & Spiers, 2002).

The 1980s saw the first main wave of qualitative literature and the emergence of a new language for research. The introduction of Lincoln and Guba's (1985) ideas on trustworthiness provided an opportunity for qualitative researchers to explore new ways of expressing validity, reliability, and generalizability outside the semantics of the quantitative paradigm (Guba & Lincoln, 1981). Lincoln and Guba recognized that their new criterion might be imperfect. It also would not yet be considered significant as it stood in marked contrast to that of quantitative inquiry, which claimed to be utterly unassailable (Tobin, 2004). Lincoln and Guba later refined the trustworthiness construct by introducing the criteria of credibility, dependability, confirmability, and transferability. These concepts were innovative and challenging, and they provided the initial platform from which much of the current debate on scientific adequacy emerged (Lincoln & Guba, 1985).

We also recognize and make consideration for the fact that Lincoln's (1985) paradigms are not only perspectives regarding scientific adequacy. As Messick (1989, 1995) thought that validity and values are one imperative, he considered that test validation implicates both the science and the ethics of assessment. As our purpose is to develop integrative framework of quantitative and qualitative criteria, we adopted the conventional definition of *validity* (in Messick's terms) in this study.

The rejection of the terms *validity* and *reliability* for evaluation came about from arguments presented in the literature emphasizing the need for a new criterion that was in no way linked to the quantitative approach (Pech, 1999; Whittemore, Chase, & Mandle, 2001). Although the reasons why terms pertaining to the quantitative paradigm that were not pertinent to qualitative inquiry have been clearly argued (Altheide, 1994; Leininger, 1994), this outright denunciation is cautioned against by Morse et al. (2002), who warned that it could result in qualitative research's being rejected as a science. Morse et al. pointed out that science is concerned with scientific adequacy and that if we reject the concepts of validity and reliability, we would reject the concept of scientific adequacy.

The extant literature highlights the increasing need to reconsider evaluation criteria for scientific adequacy in health care research and to explore an integrative framework of quantitative and qualitative criteria. In this article, we compare quantitative criteria with qualitative criteria and examine their underlying epistemologies.

Review method

As for review perspective, we first examine Patton's (1990) "paradigm of choices" (p. 39), which supports methodological appropriateness as the primary criterion for judging methodological quality. A paradigm of choices allows that different methods might be appropriate in different situations, a view that is quite different from the conventional pragmatic perspectives.

Using a paradigm of choices, in this article we explore the epistemologies underlying quantitative and qualitative evaluation criteria and discuss the issue of appropriateness between paradigm and type of study. Conducting research with an understanding of the strengths and limitations of both quantitative and qualitative methodologies is also important from the viewpoint of critical multiplism (Letourneau & Allen, 1999). To date, however, there exists no integrative framework of quantitative and qualitative criteria (Miles & Huberman, 1994). Unclear methodology can lead to a lack of scientific adequacy in research (Mays & Pope, 2001). Given the lack of an integrative framework of quantitative and qualitative criteria, we aim in this article to develop a clear definition of *evaluation criteria*. For this purpose, we review the four pairs of paradigms—validity/credibility, reliability/dependability, objectivity/ confirmability, generalizability/ transferability—previously proposed by Lincoln and Guba (1985).

A MEDLINE search was conducted to identify studies that discuss evaluation criteria for research. We excluded studies that did not mention validity, credibility, reliability, dependability, objectivity, confirmability, generalizability, or transferability if they had description regarding some evaluation criteria. We have also included articles recommended by health care research experts.

Validity of setting the research frameworks (validity/credibility)

Validity is the strength of research conclusions, inferences, or propositions. More formally, Cook and Campbell (1979) defined it as the "best available approximation to the truth or falsity of a given inference, proposition or conclusion" (p. 37). Others have explained that *validity* refers to whether the research truly measures what it intended to measure or how truthful the research results are (Joppe, 2006). The validity of quantitative research conclusions is generally divided into three categories: criterion-oriented validity, content validity, and construct validity (Carmines & Zeller, 1979; Cronbach, 1955; Nunnally, 1978; Wainer, 1988).

If a measured value and criterion score are determined at essentially the same time, a researcher can establish criterion-oriented validity. Criterion-oriented validity is also established when one test is proposed as a substitute for another or a test is shown to correlate with some contemporary criteria. Content validity is established by showing that the test items are samples of a universe in which the investigator is interested; the researcher will define a universe, and sample systematically within that universe in order to develop a framework. Establishing content validity has significant limitations, however, as it is difficult to avoid error (Bohrnstedt, 1985; Kirk & Miller, 1986). Construct validity is involved whenever a test is to be interpreted as a measure of some attribute or quality that is not operationally defined. The problem faced by the investigator is, What constructs account for variance in test performance? Seemingly, there is recognition that a research framework can be established in advance.

Alternatively, Lincoln and Guba (1985) have argued that there is actually a set of multiple constructs formed in the minds of people and that to ensure methodological scientific adequacy, researchers must demonstrate that they have recreated these multiple constructs adequately; that is, that the recreations, arrived at via inquiry, are credible to the creators of the original multiple realities. In this view, *credible* is the operational word, and credibility is enhanced by activities that make it more likely that credible findings and interpretations will be produced, such as prolonged engagement and triangulation, and activities aimed at refining a working hypotheses as more and more information becomes available (negative case analysis) (Janesick, 2000; Lincoln, 1985; Schwandt , 2001). In the case of credibility as an evaluation criterion, a research framework is formed in the process of the research, suggesting that one could not be established in advance.

As for the validity/credibility paradigm, there is a major difference regarding the creation of research frameworks (Table 1). When a research framework can be established in advance, it is used to make observations. Because, in this case, developing a research framework means conducting research from a certain perspective, it is broader than testing a hypothesis. When a research framework cannot be established in advance, it is formed in the process of the study. Although quantitative studies employ numbers, which allow limited interpretation, most qualitative studies use words, which might allow for wider interpretation. Thus, quantitative studies are well suited to validity as an evaluation criterion, whereas qualitative studies are well suited to credibility.

In some cases, however, qualitative studies, such as participant observations, do test hypotheses (Stake, 1978). In these instances it might be valuable to apply the validity criterion, as a research framework is formed in advance. Additionally, if the research objective is to form a new hypothesis, researchers might create only a partial framework in advance to, for example, prepare an interview protocol or interview structure. In this case, both credibility and validity might be appropriate criteria for evaluation.

The same is equally true for quantitative research. It is not always possible to create a rigorous framework in advance of every quantitative study. For example, some statistical analyses, such as exploratory factor analysis or data mining, allow room for interpretation, and the use of credibility as an evaluation criterion can contribute to the robustness of the research. However, current techniques of enhancing credibility are limited to text-based data, and new techniques are needed for application to numerical data. The use of questionnaires in quantitative research is another case where the use of credibility might be appropriate. The objective of a text-based questionnaire is to capture reality; therefore, there might be a strong need to use the credibility criterion for evaluation. Because using credibility as an evaluation criterion for quantitative research is useful, further exploration of this use is necessary.

The validity/credibility paradigm is centered on the establishment of a research framework. Validity is used to evaluate frameworks that are set in advance. Credibility is used to evaluate frameworks that are created in the process of research. After assessing the nature of the research framework, a researcher should use either validity or credibility, or both, as appropriate.

Stability of the phenomena and methods (reliability/dependability)

Reliability refers to the extent to which results are consistent over time and are an accurate representation of the total population under study (Nunnally, 1978). In other words, if the results of a study can be reproduced under a similar method, then the research instrument is considered to be reliable. It is important to note, however, that as each type of measurement has a certain level of nonsystematic error, it is impossible to remove systematic error completely. Kirk and Miller (1986) identified three types of reliability pertaining to quantitative research: quixotic reliability, synchronic reliability, and diachronic reliability. *Quixotic reliability* refers to the circumstances in which a single method of observation continually yields an erroneous result, which can be detrimentally deceptive. *Diachronic reliability* refers to the stability of an observation through time. In the social sciences this concept is manifested in test-retest paradigms of experimental psychology and survey research. *Synchronic reliability* refers to the similarity of observations within the same time period. It involves observations that are consistent with respect to the particular features of interest to the observer and can be evaluated by comparisons of data elicited by alternative forms (e.g., split-half testing, interrater correlation) (Cronbach, 1951; Nunnally, 1978). Kirk and Miller believed that diachronic reliability and synchronic reliability cannot always be applied to qualitative research. The general applicability of diachronic reliability is somewhat diminished by the fact that it is appropriate only for the measurement of features and entities that remain unchanged., In effect, reliability is an evaluation criterion centered on the stability of results and the recognition that phenomena and methods can assume stability in the context of the research.

Table 1. Paradigm of criteria and underlying epistemology

Category	*Epistemology*	*Paradigm of Criteria*	*Tactics*	*How to Use Criteria*
Validity of setting the research frameworks (validity)	When research frameworks could be set in advance, *they are used to make observations.*	Validity	Factor analysis, examine an external criterion, pretest	Validity is used to evaluate frameworks that are set in advance Credibility is used to evaluate frameworks that are created in the process of the research
	When research frameworks could not be set in advance, *they are formed in the process of the research When research frameworks could not be set in advance,*	Credibility	Triangulation, member check, negative case analysis	
Stability of the phenomena and methods (stability)	Phenomena and methods that could assume stability *are evaluated by the stability of results.*	Reliability	Test-retest, interrater consistency, Cronbach's alpha	Reliability is used to evaluate phenomena and methods that can assume stability Dependability is used to evaluate those that cannot assume stability
	Phenomena and methods that could not assume stability *are evaluated by the stability of research processes.*	Dependability	Consistency on data collection and analysis, warranty of the traceability, data auditing	
Neutrality of the observations and interventions (neutrality)	When observations and interventions are conducted in a neutral way, *errors made in the process of research need to be checked.*	Objectivity	Making verbatim record, conduct double-check for data entry	Objectivity is used to evaluate observations and interventions conducted in a neutral way Confirmability is used to evaluate those in which neutrality could not be secured
	When observations or interventions are not conducted in a neutral way, *their effects on processes and consequences of the research need to be checked.*	Confirmability	Making reflexive journals, using rich data, developing a rapport	
Range and applicability of findings (applicability)	As for the field where the study findings could be applied, *some examination toward generalization is needed.*	Generalizability	Random sampling, randomization, matching, comparing with parent population	After setting the range of application, generalizability is used to evaluate generalization and transferability is used to evaluate extrapolation
	As for the field where the study findings could not be applied, *some information toward extrapolation is needed.*	Transferability	Thick description of research context	

Dependability is used as an evaluation criterion when an observed phenomenon is likely to change depending on a research method, time, or environment and it is difficult to assume stability in the research (Lincoln & Guba, 1985). As a criterion dependability takes into account both factors related to instability and factors related to phenomenal or design-induced changes (Schwandt, 2001). In research studies where dependability is used as the evaluation criterion, it is often the case that researchers have to certify that the collection and analysis of data fall within acceptable professional, legal, and ethical limits (Schwandt, 2001). Dependability is assessed mainly by the use of a dependability audit, in which a third-party auditor systematically checks the researcher's work through review of the audit trail (Schwandt, 1997). The audit trail includes tape recordings, transcripts, interviewer's guides, data reduction and analysis products, and the list of working hypotheses. The auditor ascertains fairness in the representation of the research process and determines whether researchers' accounts are supported by corroborative documents. Dependability is an evaluation criterion focused on the consistency of the research process and is applicable in cases where both method and phenomena might prove to be unstable.

The stability of phenomena and methods varies greatly in the reliability/dependability paradigm (see Table 1). Phenomena and methods that can assume stability are evaluated by the stability of results, whereas phenomena and methods that are not stable are evaluated by the consistency of the research process. Reliability as an evaluation criterion is highly suited to studies with controlled research environments, such as a laboratory observation, whereas dependability is more applicable to studies measuring less controllable events, such as those dealing with human emotion.

In quantitative research the stability of phenomena and methods is often the case; however, this cannot be assumed automatically. Similarly, there are instances when the method and phenomena are actually stable in quantitative studies. Mental health research is a prime example. Symptoms of mental disorders often differ greatly across a group of patients; therefore, assessing clinical condition as a variable is not stable (American Psychiatric Association, 1994). To ensure dependability between symptoms and a diagnosis, the researcher must define a mental disorder based on diagnostic criteria. After defining the diagnosis of a mental disorder, the effects of treatment can be evaluated quantitatively and the characteristics of symptoms can be examined qualitatively. This is also an example where the dependability criterion is used to evaluate the scientific adequacy of both quantitative and qualitative results.

In health care research, rather than choosing a single evaluation criterion based on the assumption of stability (or instability), it is necessary to assess stability in terms of observational method, observed phenomenon, and data analysis to use appropriate criterion. In this paradigm reliability is used to evaluate phenomena and methods that assume stability, whereas dependability is used to evaluate those that do not assume stability.

Neutrality of the observations and interventions (objectivity/confirmability)

Regarding the distance between the observer and the observed, two perspectives are of principal importance: The first is based on the notion that a short distance harms objectivity, and the second is based on the notion that a long distance causes a lack of understanding of the observed.[1] Quantitative research generally chooses the former perspective, keeping a certain distance from the observed. Objectivity assumes three things: (a) there is an isomorphism between the study data and reality, (b) observers can keep adequate distance from the observed, and (c) inquiry is value free (Lincoln & Guba, 1985). When observations and interventions are conducted in a neutral way, human errors made in the process of research need to be vetted to ensure scientific adequacy.

Conversely, qualitative research sometimes addresses issues that require researchers to have direct emotional involvement or to experience sympathy, such as when conducting a one-on-one survey and

having to developing a cordial relationship with study participants (Lofland, 1971). In such cases, observations and interventions have an effect not only on the observed but also on the observers. When the three assumptions regarding objectivity cannot be held true, researchers need to establish confirmability by controlling for the effects of observations and interventions on the process and consequences of the research.(Lincoln & Guba, 1985) This is done by certifying that the findings and interpretations are based on raw data and by making transparent the methods and process of the research (e.g., raw data, data reduction and analysis products, data reconstruction and synthesis products, process notes) (Miles & Huberman, 1994).

In the objectivity/confirmability paradigm the difference in evaluation methods lies in the recognition of the neutrality of observations and interventions (see Table 1). When observations and interventions are conducted in a neutral way, errors made in the process of research need to be controlled for, and when observations or interventions are not conducted in a neutral way, their effects on the process and consequences of the research need to be controlled for.

In health care research the majority of participants are human beings, and research often has an effect on the observed, even in quantitative studies. Interviews, for example, have various effects on both the observers and the observed due to interpersonal contact, and in survey studies the wording, construction, and arrangement of questions in questionnaires might influence participants' answers. Other factors that can have a confounding effect on results include the researchers themselves, who could realize academic achievement by reporting the research results, and the research sponsor, who might have direct interests in the research results. Although a limited number of scientific journals currently require authors to provide information regarding the source of funding, many journals seem to lack awareness of findings that might be affected by the sponsors. Thus, it might be useful to further examine the possible influence from a sponsor or parent organization on research findings.

Range and applicability of findings (generalizability/transferability)

Generalizability is defined as the extent to which the findings obtained on a specific sample can be applied to the target population (Rothman & Greenland, 1998). This definition does not imply that all the characteristics of the sample should be similar to those of the target population, although it is intuitive that a lack of representativeness in the study sample will limit generalizability. Cronbach (1975) argued that the concept of generalizability in social sciences is different from that of the natural sciences and that researchers should give attention to both controlled and uncontrolled variables in social sciences. As researchers move from one situation to another, their first task is to describe and interpret the effect anew in each situation. If proper weight is given to local conditions, Cronbach believed generalization to be a working hypothesis, not a conclusion. When we are considering generalization, it is important to determine the range of application for research findings.

Shadish (1995) argued that the core principles of generalization apply to both quantitative and qualitative research. Findings from either type of study can be generalized according to five principles: (a) the principle of proximal similarity, (b) the principle of heterogeneity of irrelevancies, (c) the principle of discriminant validity, (d) the principle of empirical interpolation and extrapolation, and (e) the principle of explanation. If researchers recognize that certain study findings can be applied to another context, then exploration of generalization is useful. According to the generalizability paradigm, researchers need to determine the range of application of findings to evaluate the generalizability of those findings.

Alternatively, some qualitative researchers have argued that, at best, only working hypotheses may be abstracted and that the transferability of these hypotheses is determined empirically, depending on the degree of similarity between sending and receiving contexts. Donmoyer (1990) argued that researchers

should reject the conventional paradigm of generalizability in qualitative research. Researchers need to learn about both sending and receiving contexts to ensure the transferability of one's inference (Lincoln, 1985). The original inquirer cannot specify the sites to which transferability might be sought, but the implementers can. The responsibility of the original investigator is to enable someone interested in making a transfer to reach a conclusion about whether the transfer can be contemplated as a possibility. To improve the quality of transferability, original researchers are responsible for providing sufficient descriptive data for implementers to make better transferability judgments. Researchers should suggest possible methods of verification. As mentioned above, when researchers recognize that it could be difficult to generalize certain research findings, they should evaluate the possibility of extrapolation based on the transferability paradigm.

The contrast in the generalizability/transferability paradigm lies in the difference in the range of application (see Table 1). In a context where study findings can be applied, evaluation of generalizability is needed. In a context where study findings are not directly applicable, information regarding extrapolation is needed. Compared to the situation in basic research in biology or physics, it is difficult for applied science to attain generalizability, mainly because of its vulnerability to social contexts and regional environments, and the characteristics of participants. Evaluating the quality of research solely from the viewpoint of generalizability should be avoided, as the findings of applied science might be more easily useful to the direct return of the research results to the participants and society.

If a pilot intervention study is conducted for district "A," the findings need to be generalizable at least to the whole of district A, even though it might be difficult to simply apply the program findings to other countries that have different social systems and characteristics of residents. In such an instance, it might be useful to discuss not only the limitations to generalizability but also the transferability to another context. In quantitative research, it is imperative to establish techniques of transferability.

Although it might not be useful to evaluate the generalizability of certain types of qualitative research that report detailed findings on single cases, in studies using the grounded theory approach, one of the qualitative methods often intended to form a middle-ranged theory, some techniques regarding generalization for a limited range have been proposed. Even in qualitative research it is useful to evaluate the generalizability of findings after determining the range of application.

The generalizability/transferability paradigm is concerned with the range of application of findings. Whether it is quantitative or qualitative research, it is important to recognize and identify the range of application of research findings. After the range of application has been set, generalizability is used to evaluate generalization and transferability is used to evaluate extrapolation. Unless research findings could assume universal generalization or have no possibility of extrapolation, it is useful to use both paradigms for evaluation.

Conclusion

In this article we examined the epistemological issues behind evaluation criteria for scientific adequacy in research. Comparing quantitative research paradigms with those of qualitative research, we discussed the possible use of evaluation criteria based on a pragmatic perspective. Validity/credibility is concerned with research framework, whereas reliability/dependability refers to the stability of observations, neutrality/ confirmability reflects influences between observers and subjects, and generalizability/transferability are epistemologically different in the way findings are applied. Qualitative studies, however, do not always choose qualitative paradigms; and if we can assume stability to some extent in a qualitative study, it might be better to use a quantitative paradigm (reliability). Similarly, it is useful to employ qualitative paradigms to enhance the scientific adequacy of a quantitative study that did not use a research framework with stability in all phases of observation. Regardless of whether it is quantitative or

qualitative research, it is important to recognize the four epistemological axes. Evaluating scientific adequacy in a study should be done after establishing a framework(s), assessing stability of the phenomena and methods, ensuring neutrality of the observations and interventions, and determining the range of application for the findings.

References

Abell, P. (1990). *Methodological achievements in sociology over the past few decades with special reference to the interplay of qualitative and quantitative methods*. London: Macmillan.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alkin, M. C. (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Altheide, D. J. (1994). *Criteria for assessing interpretive validity in qualitative research*. London: Sage.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barbour, R. (1999). The case for combining qualitative and quantitative approaches in health services research. *Journal of Health Services Research and Policy, 4*, 39–43.

Bernstein, I., & Freeman, H. E. (1975). *Academic and entrepreneurial research: Consequences of diversity in federal evaluation studies*. New York: Russell Sage.

Bohrnstedt, G. W. (1985). *Measurement*. New York: Academic Press.

Bravata, D. M., McDonald, K. M., Shojania, K. G., Sundaram, V., & Owens D. K. (2005). Challenges in systematic reviews: Synthesis of topics related to the delivery, organization, and financing of healthcare. *Annals of Internal Medicine, 142*, 1056–1065.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Cook, T., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L. J. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.

Donmoyer, R. (1990). *Generalizeability and the single case study*. New York: Teachers College Press.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco: Jossey-Bass.

Hammersley, M. (1992). *Deconstructing the qualitative-quantitative divide*. Aldershot, UK: Avebury.

Janesick, V. J. (2000). *The choreography of qualitative research design: Minuets, improvisations and crystallization*. Thousand Oaks, CA: Sage.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*, 7, 14–26.

Joppe M. (2006). *The research process*. Retrieved February 10, 2009, from http://www.uoguelph.ca/htm/MJResearch/ResearchProcess/default.html

Kirk, J. M., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. Beverly Hills, CA: Sage.

Klave S. (1995). The social construct of validity. *Qualitative Inquiry, 1*, 19–40.

Koch, T. (1994). Establishing rigor in qualitative research: The decision trail. *Journal of Advanced Nursing, 19*, 976–986.

Kuhn, T. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Leininger, M. (1994). Evaluation criteria and critique of qualitative studies. In J. Morse (Ed.), *Critical issues in qualitative research methods* (pp. 95–115). Newbury Park, CA: Sage.

Letourneau, N., & Allen, M. (1999). Post-positivistic critical multiplism: A beginning dialogue. *Journal of Advanced Nursing, 30*, 623–630.

Lincoln, Y. S. (1985). Emerging criteria for qualitative and interpretive research. *Qualitative Inquiry, 3*, 275–289.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage.

Lofland, J. (1971). *Analyzing social settings*. Belmont, CA: Wadsworth.

Mays, N., & Pope, C. P. (2001). *Quality in qualitative health research*. .London: BMJ.

Mechanic, D. (1989). Medical sociology: Some tensions among theory, method and substance. *Journal of Health and Social Behavior, 30*, 147–160.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(*2), 5–11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9) 731–739.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

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), Article2. Retrieved January 28, 2009, from https://ejournals.library.ualberta.ca/index.php/IJQM/issue/archive

Mulrow C., Langhorne P., & Grimshaw, J. (1997). Integrating heterogeneous pieces of evidence in systematic reviews. *Annals of Internal Medicine, 127*(11), 989–995.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Byrne, P. (2007). The advantages and disadvantages of mixing methods: An analysis of combining traditional and autoethnographic approaches. *Qualitative Health Research, 17*(10), 1381–1391.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park, CA: Sage.

Pearlin L. (1992). Structure and meaning in medical sociology. *Journal of Health and Social Behavior, 33*, 1–9.

Pech, E. S. (1999). Quality criteria for qualitative research: Does context make a difference? *Qualitative Health Research, 9*, 552–558.

Pope, C., & Mays, N. (2001). *Qualitative methods in health research*. London: BMJ.

Rothman, K. J., & Greenland, S. (1998). Precision and validity in epidemiologic studies. In K. J. Rothman & S. Greenland (Eds.), *Modern epidemiology* Philadelphia: Lippincott, Williams & Wilkins.

Schwandt, T. A. (1997). *Qualitative inquiry: A dictionary of terms*. Thousand Oaks, CA: Sage.

Schwandt, T. A. (2001). *Dictionary of qualitative inquiry*. Thousand Oaks, CA: Sage.

Shadish, W. R. (1995). Philosophy of science and the quantitative-qualitative debates: Thirteen common errors. *Evaluation & Program Planning, 18*, 63–75.

Stake, R. E. (1978). The case study method in social inquiry. *Educational Researcher, 7*(2), 5–8.

Thomas S. (1996). The virtue of qualitative and quantitative research. *Annals of Internal Medicine, 125*, 770–771.

Tobin, G. A. (2004). Methodological rigour within a qualitative framework. *Journal of Advanced Nursing, 48*, 388–396.

Watson, L. A., & Girard, F. M. (2004). Establishing integrity and avoiding methodological misunderstanding. *Qualitative Health Research, 14*(6), 875–881.

Wainer, H. B. (1988). *Test validity*. Hilldale, NJ: Lawrence Erlbaum.

Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research, 4*, 522–537.