The Holy Trinity of Methodological Rigor: A Skeptical View

Chris L. S. Coryn
The Evaluation Center—Western Michigan University

> Methodologists are the preachers of science. Armed with canons of correct procedure, they have the power to castigate and exhort. They can instruct us to have clearly defined objectives and explicit frames of reference, to base our studies on good theories…the process of science does not work from rules to practice but from attempt to attempt…one good piece of research influences research practice more than many methodology textbooks (Przeworski, 1987, p. 31).

Rigor in research is normally conceived of as the means by which integrity and competence are confirmed (Tobin & Begley, 2004).[1] That is, a way of demonstrating the legitimacy or soundness of the research process. Without rigor, it is argued, there is a danger that research may become fictional journalism and therefore worthless as contributing to knowledge (Morse, Barrett, Mayan, Olson, & Spiers, 2002).

A substantial proportion of the scientific community associates research quality with methodological rigor (Farrington, 2003), though it only constitutes a small segment of the scientific method.[2] However, since its introduction by Donald Campbell and Julian Stanley in the early 1960s, followed by its successors in 1979 (Cook & Campbell) and 2002 (Shadish, Cook, & Campbell), hierarchies of evidence, as they have come to be known in some circles, often form the basis by which research quality is judged.[3] Foremost amongst users of these types of hierarchies are the health sciences (Boaz & Ashby, 2003; Grayson, 2002). The hierarchy of evidence often employed to judge methodological rigor, or soundness, especially of quantitative research is (in descending order from highest to lowest quality, or rigor):[4]

1. Systematic reviews and meta-analyses[5]
2. Well-designed randomized controlled trials

[1] Rigor literally means 'the quality of being unyielding or inflexible' or 'strict precision.'

[2] Many undergraduate and graduate program now include courses intended to make them good consumers and evaluators of research using textbooks, such as *Evaluating Research Articles* (Girden, 2001), which is intended to train students in evaluating the soundness of research designs and appropriateness of statistical analyses in the published research literature.

[3] This does not imply that standards of methodological rigor did not exist prior to Campbell and Stanley's *Experimental and Quasi-Experimental Designs for Research* (1963), only that these have become the standards by which much research has been upheld in the last four decades.

[4] There are many variations of hierarchies related to methodological rigor. The one presented here is primarily applied to address the effectiveness of clinical interventions.

[5] The placing of meta-analysis and systematic reviews on the hierarchal ladder, let alone at the top, is interesting given that there are incredible variations in the quality of the two, and they are not necessarily viewed as research designs per se, but as methods to review literature.

3. Well-designed trials without controls (e.g., single-group pre-post, time series or matched case-controlled studies)
4. Well-designed non-experimental studies from more than one center
5. Opinion of respected authorities, based on clinical evidence, descriptive studies or reports of expert committees

These hierarchies are grounded in the quantitative tradition and usually consist of the criteria of validity, reliability, and objectivity. Combined, these criteria have almost reached the status of a 'holy trinity' (Kvale, 1995; Tobin & Begley, 2004).[6] Yet, the exact nature of validity has eluded adequate, and agreed upon, characterization since there exists no single or common explanation of the term. Common definitions include "[research is valid]…if it represents accurately those features of the phenomena that it is intended to describe, explain or theorise" (Hammersley, 1987, p. 69) and "the truth of, correctness of, or degree of support for an inference" (Shadish, Cook, & Campbell, 2002, p. 513).[7, 8] The traditional view of reliability, on the other hand, is premised on assumptions of replicability, repeatability, or consistency (Golafshani, 2003; Shadish, Cook, & Campbell, 2002).[9] However, reliability is often portrayed as the extent to which a measurement or observation yields the same answer or results however and whenever it is carried out. As a pair, validity and reliability are sometimes described as two complementary aspects of objectivity (Tobin & Begley, 2004) to the extent that objectivity is normally understood a being free from bias (e.g., cognitive, cultural, sampling) or distortion (Trochim, 2002).[10]

Debate around the relevance and use of this and other versions of the trinity as standards of research quality, or methodological rigor, has persisted for more than 20 years (Flick, 2006; Guba & Lincoln, 1981; Morse, Barrett, Mayan, Olson, & Spiers, 2002; Sandelowski, 1986; Tobin & Begley, 2004; Winter, 2000; Yin, 1994), and Guba (1981) warned that these criteria are "primitive" (p. 90), and should be applied only as guides rather than orthodoxy.

By far, the most energy has been devoted to developing standards for assessing the quality of randomized controlled trials (RCTs), which are usually:[11]

1. Was the assignment to the treatment groups really random?

[6] Other versions of the trinity often include generalization rather than objectivity as a criterion for judging the soundness, rigor, or quality of research (Kvale, 1995). However, internal and external validity are usually considered the essential elements of generalization, especially the latter. For instance, statistical sampling is often the dominant basis for generalizing (Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2004; Kish, 1965).

[7] In fact, there is presently no 'index' of validity.

[8] Or, "…an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (Messick, 1989, p. 13).

[9] In order to estimate reliability, researchers often apply various theoretical or statistical assumptions such as true score theory (Spearman, 1907, 1913) or generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). However, reliability in terms of design should not be confused with reliability from the test theory notions. The reliability criterion of methodological rigor is more about replicability, repeatability, or consistency, whereas in test theory it is more a statement about common variance shared.

[10] However, in science, the ideal of objectivity is generally considered to come about as a result of strict observance of the scientific method. Objectivity in science is intimately related to the aim of reproducibility. Methodological aspects can be roughly distinguished as objectivity in measurement and objectivity in experimentation and interpretation. As such, it is only tangentially related to the concept of objectivity in philosophy, and closer to, for example, objectivity in journalism. Another methodological aspect is the avoidance of bias, which can involve cognitive bias and cultural bias, but also sampling bias. Methods for avoiding or overcoming such bias include random sampling and double-blind trials.

[11] Randomization purports to control an infinite number of rival hypotheses without ever specifying what any of them are, and when RCTs are compromised there is a tendency for the results of systematic reviews and meta-analyses to be distorted (Jüni, Altman, & Egger, 2002).

2. Was the treatment allocation concealed?
3. Were the groups similar at baseline in terms of prognostic factors?
4. Were the eligibility criteria specified?
5. Were outcome assessors blinded to the treatment allocation?[12]
6. Was the care provider blinded?
7. Was the patient blinded?
8. Were the point estimates and measures of variability presented for the primary outcome measure?
9. Did the analyses include an intention to treat analysis?

In part, these standards have been the result of social scientists trying to replicate the natural sciences, such as physics and chemistry, as the ideal embodiments of scientific inquiry (Scriven, 2006). Nowhere is this dogmatic view of methodological rigor more apparent than the United States Department of Education's Institute for Educational Science (IES) which allocates and controls a budget of nearly $500 million and is now only funding RCTs (Donaldson & Christie, 2005; Scriven, 2006, September).

Criteria for assessing the quality of qualitative research have also emerged in order to address the 'fitness for purpose' of research, although, their application often involves some redefinition of the terms:[13]

> The usual canons of good science have value but require redefinition to fit the realities of qualitative research and the complexities of the social phenomena that we seek to understand (Strauss & Corbin, 1998, p. 266).

Like the quantitative tradition, most criteria for assessing qualitative research have come from the field of health studies (Boulton & Fitzpatrick, 1994; Lincoln, 1992; Lincoln & Guba, 1990; Mays & Pope, 1995). One of the most widely applied lists comes from the Medical Sociology Group (1996), which includes the following criteria for assessing the soundness of qualitative research:

1. Are the research methods appropriate to the question being asked?
2. Is there a clear connection to an existing body of knowledge?
3. Are the criteria for/approach to sample selection, data collection and analysis clear and systematically applied?
4. Is the relationship between the researchers and researched considered, and have the latter been fully informed?
5. Is sufficient consideration given to how findings are derived from the data and how the validity of the findings were tested?
6. Has evidence for and against the researcher's interpretation been considered?
7. Is the context for the research adequately described and accounted for?
8. Are findings systematically reported and is sufficient original evidence reported to justify a relationship between evidence and conclusions?
9. Are researchers clear about their own position in relation to the research topic?

From this and similar lists, the qualitative tradition ultimately developed its own holy trinity, which includes the criterion of trustworthiness (including subcriteria of credibility and transferability), the criterion of dependability, and the criterion of confirmability (Flick, 2006; Golafshani, 2003; Guba & Lincoln, 1989; Trochim, 2002).[14, 15]

[12] Blinding is sometimes referred to as masking (Shadish, Cook, & Campbell, 2002).

[13] Fitness for purpose is the notion that in assessing the quality of research that methodological rigor is but one aspect of quality; the other aspect is the relevance of the research for policy or practice (Boaz & Ashby, 2003; Patton, 2002).

[14] Guba and Lincoln (1989) felt that their four criteria better reflected the underlying assumptions involved in most qualitative research.

[15] Flick (2006), for example, discusses qualitative validity in terms of "whether researchers see what they think they see" (p. 371). Moreover, he and others (Kirk & Miller, 1986) argue that three types of error may occur as regards qualitative validity: seeing a relationship, a principle, and so on when they are not correct (Type I error); to reject them when they are correct (Type II error); and asking the wrong questions (Type III error). Type III error

Trustworthiness is the extent to which the results of qualitative research are credible or believable from the perspective of the participant in the research (i.e., credibility) and the degree to which results can be generalized or transferred to other contexts or settings (i.e., transferability). The dependability criterion emphasizes the need for the researcher to account for the ever-changing context within which research occurs as most qualitative researchers tend to assume that each researcher brings a unique perspective to the study. Confirmability refers to the degree to which the results can be confirmed or corroborated by others (Trochim, 2002).

To the extent that the generally accepted quantitative and qualitative criteria of methodological soundness, or quality, differ is subject to dispute, and the similarities are strong enough that it can be reasonably inferred that they are in fact comparable. For instance, the qualitative subcriteria of the trustworthiness criterion (i.e., credibility and transferability) are simply parallels of the quantitative concepts of internal and external validity, where credibility is synonymous with internal validity and transferability is congruent to external validity. The dependability criterion, on the other hand, is analogous in most respects to the quantitative criterion of reliability and confirmability is in essence the quantitative standard of objectivity. Although the quantitative and qualitative trinities represent 'aspects' of good research, in and of themselves, they are not sufficient to judge the merits of an instance of research, and certainly not its worth or significance.

In any case, rigor in terms of integrity, competence, legitimacy, or soundness is one of sufficiency. The correct standard, or basis, in science or outside it, for such conclusions is that they can be demonstrated or established *beyond reasonable doubt* (Scriven, 2006) and this is crucial for evaluating certain aspects or properties of 'good' research (Coryn, 2006, 2007).[16]

occurs when the right answer to the wrong question is observed (Raiffa, 1968).

References

Boaz, A., & Ashby, D. (2003). *Fit for purpose? Assessing research quality for evidence based policy and practice*. London, England: University of London, ESRC UK Centre for Evidence Based Policy and Practice.

Boulton, M., & Fitzpatrick, R. (1994). 'Quality' in qualitative research. *Critical Public Health*, *5*(3), 19-26.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand-McNally.

Coryn, C. L. S. (2006). The fundamental characteristics of research. *Journal of Multidisciplinary Evaluation*, *5*, 124-133.

Coryn, C. L. S. (2007). *Evaluation of researchers and their research: toward making the implicit explicit*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurements. New York, NY: John Wiley & Sons.

Donaldson, S. I., & Christie, C. A. (2005). The 2004 Claremont debate: Lipsey versus Scriven—Determining causality in program evaluation and applied research: Should experimental evidence be the gold standard? *Journal of MultiDisciplinary Evaluation*, *3*, 60-77.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science*, *587*(1), 49-68.

[16] Beyond reasonable doubt, "the standard of evidence that the courts require in felony trials, is far stronger than 'the balance of evidence' the standard they use, and sharply distinguish, for misdemeanor trials. People sometimes think that RCTs are the paradigm design because they meet some higher standard than beyond reasonable doubt, perhaps 'beyond the practical possibility of error.' But they are far from that standard, which is not even met by proofs in deductive logic and mathematics" (Scriven, 2006, p. 4)

Flick, U. (2006). *An introduction to qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Girden, E. R. (2001). *Evaluating research articles* (2nd ed.). Thousand Oaks, CA: Sage.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597-607.

Grayson, L. (2002). *Evidence based policy and the quality of evidence: rethinking peer review* (Working Paper #7). London, England: University of London, Department of Politics, Centre for Evidence Based Policy and Practice.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal*, *29*(2), 75-91.

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

Hammersley, M. (1987). Some notes on the terms 'validity' and 'reliability'. *British Educational Research Journal*, *13*(1), 73-81.

Jüni, P., Altman, D. G., & Egger, M. (2002). Assessing the quality of randomized controlled trials. In D. G. Altman, M. Egger, & G. D. Smith (Eds.), *Systematic reviews in healthcare: Meta-analysis in context* (pp. 23-42). London, England: BMJ Books.

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

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley & Sons.

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

Lincoln, Y. S. (1992). Sympathetic connections between qualitative methods and health research. *Qualitative Health Research*, *2*(4), 375-391.

Lincoln, Y. S. (1995, April). *Emerging criteria for quality in qualitative and interpretive research*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Lincoln, Y. S., & Guba, E. G. (1990). Judging the quality of case study reports. *Qualitative Studies in Education*, *3*(1), 53-59.

Mays, N., & Pope, C. (1995). Qualitative research: Rigour and qualitative research. *BMJ*, *311*, 109-112.

Medical Sociology Group. (1996). Criteria for the evaluation of qualitative research papers. *Medical Sociology News*, *22*(1), 68-71.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

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), Article 2. Retrieved July 16, 2006 from http://www.ualberta.ca/~ijqm/

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Przeworski, A. (1987). Methods of cross-national research, 1970–1983: An overview. In M. Dierkes, H. N. Weiler, & A. B. Antal (Eds.), *Comparative policy research: Learning from experience* (pp. 31-49). Aldershot, England: Gower.

Raiffa, H. (1968). *Decision analysis*. Reading, MA: Addison-Wesley.

Sandelowski, M. (1995). The problem of rigor in qualitative research. *Advances in Nursing Science*, *8*(3), 125-130.

Scriven, M. (2006). *Can we infer causation from cross-sectional data?* Retrieved August 31, 2006 from http://www7.nationalacademies.org/bota/School-Level Data_Michael Scriven-Paper.pdf

Scriven, M. (2006, September). *The latest battle in the war over research designs for establishing*

causation. Paper presented at The Evaluation Center's Evaluation Café series, Kalamazoo, Michigan.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Spearman, C. (1907). Demonstration of the formulae for true measurement of correlation. *American Journal of Psychology*, *18*(2), 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology*, *5*, 417-426.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Tobin, G. A., & Begley, C. M. (2004). Methodological rigour within a qualitative framework. *Journal of Advance Nursing*, *48*(4), 388-396.

Trochim, W. M. K. (2002). *Qualitative validity*. Retrieved July 18, 2006 from http://www.socialresearchmethods.net/kb/qualval.htm

Winter, G. (2000). A comparative discussion of the notion of 'validity' in qualitative and quantitative research. *The Qualitative Report*, *4*(3/4). Retrieved July 16, 2006 from http://www.nova.edu/ssss/QR/QR4-3/winter.html

Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed.) (Applied Social Research Methods Series, Vol. 5). Thousand Oaks, CA: Sage.