Guest Editorial: Educational Research as Disciplined Inquiry: Examining the Facets of Rigor in Our Work

Thou shalt not write thy doctor's thesis on education[1]

These words are from "Under Which Lyre" by the British poet Wystan Auden (1907–1973) in his Phi Delta Kappa Poem at Harvard in 1946, shortly after the end of the Second World War. Auden's poem, subtitled "A Reactionary Tract for the Times," rails against elements of a mass tertiary education and celebrates both intellectual and sensual exploration. His concerns about first-year students at Harvard, Princeton, and Yale remind me of teacher education students who approach their programs with the view that there is a finite set of clear rules of procedure that, when followed, lead to good teaching. Research students mirror this behavior if they slavishly follow detailed steps for ensuring that their quasi-experimental designs meet criteria for validity. Neither are students of qualitative research immune from the infection of checklists. As Webb and Glesne (1992) noted, "Some students assume that a qualitative research class will provide procedures that, if followed faithfully, will produce warranted research results" (p. 775). But what is meant by "warranted research results"? For me there is more to "warranted research results" than having the researcher satisfy familiar checklists, such as those in the five editions of McMillan and Schumacher's *Research in Education: A Conceptual Introduction* (McMillan & Schumacher, 1984–2001)—a text that I use once or twice a year when I teach an introductory research methods course. In this course, I try to have students understand that, ultimately, our research is a human enterprise and that its worth is more than simply its trustworthiness. So although "warranted research results" has something to do with trustworthiness and with concepts like reliability and validity, I know that there is more. In this piece, I explore the character of rigor and its place in discussions about the quality of our research, so that we can move away from overworked paradigm wars and from the ambiguity of reliability and validity.

There was a time when the idea that rigor might reside in qualitative educational research was almost ridiculed. But paradigm wars existed in other disciplines before we declared them in education, so issues of rigor lie beyond debates about qualitative and quantitative research. For example, among my readings as a doctoral student in the late 1960s was the debate about the function of history with Hempel (1968) proposing explanations for human behaviors as

An earlier version of this paper was an invited address at the annual meeting of the National Association for Research on Science Teaching, St. Louis, MO, March 2001. This paper is from the research program "Co-op Education and Workplace Learning" (Hugh Munby, Nancy Hutchinson, and Peter Chin, Principal Investigators).

This research is funded by the Social Sciences and Humanities Research Council of Canada.

[1] The study of education was not the only object of Auden's witty scorn (Auden, 1979): "Do not sit with statisticians nor commit a social science" is an equally memorable injunction from the same poem.

DOI 10.1002/sce.10096

deductions from general laws; against this, Dray (1957) and others arguing for the unique character of historical explanation lodged within a singular context. I recently encountered this conversation continued in letters between proponents and opponents of quantification in Aydelotte's *Quantification in History* (Aydelotte, 1971), and I recalled the importance of distinguishing between two kinds of argument in this context. The first kind is about what overarching approach is proper and should be taken to research. The debate between Hempel and Dray typifies this kind of argument. The second kind of argument is about the quality of the research itself, and this is where reliability and validity have found employment.

The other side to the paradigm wars is the evidence that we were and still are comfortable when scholarly work from one paradigm is approached from the vantage point of another. An example is Conant's richly informative and literary accounts of experimental science from an historical perspective (Conant, 1957), which opened my eyes to the fundamental character of science. What I saw was quite different from what I had been taught at school and university, yet it enriched what I had been taught. So it is that I find Bruner's claim (Bruner, 1996), "The process of science making is narrative" (p. 126), unsurprising as he describes the difference between finished science and the "lively processes" (p. 127) of science making. In this light, we can imagine how we have become accepting of alternative approaches to educational research. Similarly, we should not be surprised to find sections on bivariate and multivariate statistics in LeCompte and Schensul's *Analyzing and Interpreting Ethnographic Data* (LeCompte & Schensul, 1999). It is as if "paradigm wars" have become paradigm rapprochements in which different viewpoints let us see better the human condition within our research. In turn, this suggests to me that there is more to assessing research than what is conveyed by reliability and validity.

RELIABILITY AND VALIDITY: MEANINGS FROM REPUTABLE SOURCES

The terms validity and reliability handily predate current use in research,[2] but the student of research will find precise contemporary meaning elusive. For example, reliability in the quantitative social sciences is connected with the reproducibility of results, and it has come to be associated with agreement across cases and observations. Most particularly, the term becomes a property of instruments for mental measurement (Gould & Kolb, 1964), although reliability or stability of data can be concerned with the reliability of the observer, the coder, and the analyst. And this sense seems to coincide with how the term may appear in qualitative research, especially in ethnographic work. But as we pursue this, so the matter becomes complex. For example, studies themselves, experimental or descriptive, can be judged for reliability. Goetz and LeCompte (1984) state that "Reliability refers to the extent to which studies can be replicated" (p. 211), and so "external reliability addresses the issue of whether independent researchers would discover the same phenomena or would generate the same constructs in the same or similar settings" (p. 210) while internal reliability "refers to the degree to which other researchers, given a set of previously generated constructs, would match them with data in the same way as did the original researcher" (p. 210). This becomes confusing when the authors attempt crisp definitions of validity (p. 210):

> Internal validity refers to the extent to which scientific observations and measurements are authentic representations of some reality; external validity refers to the degree to which such representations can be compared legitimately across groups.

[2] For example, the *Oxford English Dictionary* (Simpson & Weiner, 1989) reports that "validity" was used circa 1550, and that Coleridge used "reliability" in 1816 some 90 years before the term appeared in the *American Journal of Psychology*.

These ideas cohere well with the entries in *A Dictionary of the Social Sciences*, although we might be justified in being confused by the idea of "corroboration of one's data" (Gould & Kolb, 1964, p. 742) because it resembles ideas of reliability. The same sense of corroboration exists in McMillan and Schumacher's account of internal validity (McMillan & Schumacher, 1997): "Validity of qualitative designs is the degree to which the interpretations and conceptual categories have mutual meanings between the participants and the researcher" (p. 404). And their definition of external validity differs from the definition of Goetz and LeCompte (1984), but they put the concept in terms of comparability or extension, even usefulness of a study, "the degree to which the research design is adequately described so that researchers may use the study to extend the findings to other studies" (p. 411). The student might be further frustrated in a quest for clarity if he or she encountered texts in which conventional terms like "internal validity," "external validity," "reliability," and "objectivity" are replaced respectively by "credibility," "transferability," "dependability," and "confirmability" (Hoepfl, 2000, p. 9). Other versions of the terms and their meanings exist, as in Moschkovich and Brenner (2000, p. 479).

My confusion is not eased when I consider the large number of terms used in qualitative research for expressing validity and reliability, nor when I see that meanings tend to be somewhat mobile. My experience is that this state of affairs is unwelcome to graduate students in research courses, but more importantly it may be leading me and them in the wrong direction. The enterprise of discussing validity and reliability from varying viewpoints can too quickly involve us in debates about word usage. Unfortunately, this can distract us from seeing that research at its most fundamental is an *argument* that leads us through purpose, related literature, data, and analysis to a specific point.[3]

RIGOR, ARGUMENT, AND POINT

I believe we can make progress if we focus on argument itself. Certainly, the confusion about validity in mental measurement has profited from a similar switch in vantage point. Views of test validity have changed markedly over the last half century. In the 1950s, validity was construed in four separate ways: content validity, predictive validity, concurrent validity, and construct validity (APA, 1954). And 35 years later, we find Messick (1989) developing his position that these are not separate forms but are evidence for the one form: construct validity. Validity in this frame is an *argument*. A similar commitment to argument is evident in Mischler's position (Mischler, 1990). He draws on his experience in narrative research to show that validity is less important than the process of validation. He argues that "validation is the social construction of a discourse through which the results of a study come to be viewed as trustworthy for other investigators to rely upon in their own work" (p. 426). The attention to process suggests the promise of looking at argument to get a fuller sense of what is involved in the concepts of reliability and validity and how they might contribute to rigor and to showing the human character of our research.

I am far from comfortable that validity and reliability tell us all that should be told about the quality of research in education. The concepts seem necessary but not sufficient for a full account. What discussions of trustworthiness, credibility, reliability, validity seem to lack is the sense that *research has a purpose*. Here I am not referring to the purpose we might find in a section called "Statement of Purpose." Rather, I am interested in what we think research that we do is *for*: What is its point? Again, the standard accounts are a little deceptive.

[3] Roberts (1982) uses diagrams from Toulmin's ideas about the structure of argument to show that although different empirical research modes have different metaphysical premises and warrants, they share a common commitment to argument based on data.

McMillan and Schumacher (2001), in their latest edition, announce, "Research advances knowledge and improves practice" (p. 17). In fairness they then consider several different uses of research and develop these into basic (pure or fundamental), applied, and evaluative functions of research. None of this is contestable, it's just incomplete. For example, it fails to acknowledge that, among other things, research is to persuade. In the next section, I explore aspects of rigor by considering research purposes. All this is to suggest that conventional tools like reliability and validity are simply not up to the task of portraying what needs to be said about the quality and usefulness of research.

What I think we need to show to our research students is how the constructs we build in educational research get transported into arenas of professional practice, into the settings in which they can be used. My experience is that this transportation is not always successful. There seems to be a membrane between the constructs of research and arenas of practice. Presumably, if our constructs were objective, in some sense, the membrane would be easy to cross. But that option is not available to us.

> When educational researchers no longer see the possibility of objectivity as a life option, one reaction has been to focus on their subjectivity, to worry about it, and to turn it into a set of methodological concerns. For a number of researchers, anxiety about how to be as objective as possible has been translated into anxiety about how to manage subjectivity as rigorously as possible. (Heshusius, 1994, p. 15)

There are several ways in which researchers have reacted to the challenge. Heshusius, for example, advocates a methodology of participatory consciousness. My approach is rather different, indeed it starts from a different place. Basically, I do not think I have ever been wedded to objectivity itself because of the character of the knowledge produced by educational research, and because of its point. Indeed, I find a focus on point or purpose particularly helpful in describing something of the range of debates that we should enter when we consider rigor seriously. As I show below, these debates should include issues of ethics, professionalism, and rhetoric.

ETHICS AND RIGOR

I became concerned about issues of ethics and their relationship to rigor within educational research when I was asked to write on the significance of qualitative research (Munby, 1983, p. 424).

> The unquestionable purpose of the enterprise of educational research is the improvement of education. Generally, setting aside research that is more conceptual in nature, it is easy to see that quantitative and qualitative investigations of school events are designed to improve what occurs in educational institutions. While the foci of this work may run from research on classroom learning to research on curriculum change, the ultimate change held as the end-in-view has to be change in teaching practice, because what really counts is the chalkface, curriculum-in-use facet of the endeavor. Here, though, there is an implicit assumption that teaching is the sort of activity that *can* be changed. The corollary is that teachers can be changed. Of course, accompanying these assertions is the driving belief that research is worthwhile because teachers need to be changed.

For me, it is important to capture the idea that research activity begins with a normative premise. It has never been sufficient to justify research in terms of knowledge for its own sake. Indeed, I have come to think that all propositional knowledge is in the service of action (Munby, Chin, & Hutchinson, in press), and action is clearly normative. In part, the

normative nature of our research is reflected in our insistence that there be a rationale for the work. My hope is that the insistence carries into explicit statements about the value premises underlying the proposed work. Without these, I think the research would fall somewhat short of meeting a standard of rigor, and that standard is patently not an objective one when value premises are at issue.

PROFESSIONALISM AND RIGOR

In an earlier paper, I argued the impossibility of smoothly moving from generalizable research results to changing teachers principally because the particular circumstances of a teacher's action will be different from those in which the research was conducted (Munby, 1983). In quantitative research, we recognize this issue as a version of the separations among target population, sample population, and sample. In qualitative research we recognize this issue as part of the character of the research too: there is no pretense to generalizability. Here the membrane between research knowledge and professional practice is more than a matter of logic though. Professional assimilation in the field also plays a part.

Discussions about professionalism are almost as wide-ranging as definitions of what professionalism entails: in the latter we find (a) the contrast between doing something for pay and doing something free, (b) the idea that being professional involves a distancing or detachment (as in calling penalties while refereeing), (c) the suggestion that a degree of proficiency if not excellence has been achieved, and (d) the social distinction among classes that might be reflected in discriminating among occupations, vocations, and careers (Soder, 1990). Some of these discussions tend to agree that professionalism is bound to the idea of a professional knowledge base (e.g., Fenstermacher, 1990).

Colleagues and I have argued that "the essence of professionalism is professional action" (Munby, Russell, & Martin, 2001) and that teaching should be in the best interests of the clients and thus based upon the best available knowledge. But as shown in our chapter in the fourth edition of the *Handbook of Research on Teaching*, the character of teachers' knowledge is the subject of debate and conflicting theoretical viewpoints. This makes the transition of research knowledge into professional practice highly complex. And in turn, questions about the quality and value of educational research automatically get extended beyond the simple language of reliability and validity. A sense of rigor is called for that honors both the moral premises of research purpose and the integrity of professional knowledge and judgment, without violating the professionalism of the educator.

RIGOR AND THE RESEARCHER'S PROFESSIONALISM

The *Oxford English Dictionary* (Simpson & Weiner, 1989) reports many senses of "rigor" from strict application of the law, through hardness of heart, to strict accuracy and severe exactitude, a phrase that seems to refer to lexicography itself. Also we have seen how rigor gets entwined with professionalism, so it is fitting to turn the lens on the professional actions of the researcher herself or himself and to ask how rigor gets played out in that arena.

I doubt that I am alone in wondering along with graduate students at the quantity of research decisions we face that are not strictly guided by anything epistemological. Questions like, "How many participants should I really have?" "Are eight interviews enough?" "Should I attempt another administration of the test or simply go for a split-half assessment of consistency?" As a graduate supervisor, I often find myself saying, "This is just a masters thesis, not a career" thereby truncating research on purely pragmatic grounds. Of course, the section titled "Limitations" always accounts for how practicality may compromise rigor, but

somehow we miss saying that we, as professional researchers, do this all the time. Among my favorite quiet compromises are the following.

The first is the practiced convention we seem to have developed for the reliability of instruments. Noting that reliability is a function of the nature of the trait (construct) being measured, McMillan and Schumacher (2001) state, "a reliability of .80 or above is generally accepted for achievement variables, whereas estimates of .70 may be acceptable for measuring personality traits" (p. 249). I spent an early part of my career wondering about the reliability and validity of attitude measures. I won't go into details here, but it is worth observing that reliability in this sense has become something of a rhetorical device rather than an epistemological one.

The second example is the threat to internal validity of treatment replications: "In an experiment, the treatment is supposed to be repeated so that each of the members of one group receives the same treatment separately and independently of the other members of the group" (McMillan & Schumacher, 2001, p. 191). If an instructional treatment is conducted once in one class, then the class is like one subject. The sample size is the number of *treatments*, not the number of *subjects*. The threat of treatment replications refers to instances when the reported number of subjects is not the same as number of treatments.

I wonder if we are deceiving our students when we fail to show the shortcuts that we take. Of course, corners have to be cut because life is short and we cannot wait upon certainty. I am not defending compromises, but I am asking that we acknowledge that rigor is deeply connected to them in our own professional practice.

RIGOR, PERSUASION, AND RHETORIC

I know that I am not alone in trying to push for inspecting aspects of rigor in qualitative research. Sandelowski (1993) for example, recognizing "the danger of succumbing to 'the illusion of technique'" (p. 1), argued that "rigor is less about adherence to the letter of rules and procedures than it is about fidelity to the spirit of qualitative work" (p. 2). True to a certain extent, but too ephemeral for me. I think rigor refers to more than the spirit of the research, whether qualitative or quantitative. As we have seen, an element of rhetoric seems to be lurking in some of the steps we take in our research. Some argue that the element of rhetoric in quantitative research is of significant proportions:

> The language of statistics is but one form of rhetoric; however, it is a rhetoric that, for certain audiences and in certain circumstances can be more compelling and more functional than a case study, poem, or autoethnographic report. (Gergen & Gergen, 2000, p. 1033)

The term "rhetoric" may have unjustly received bad press. Although the term is sometimes used to reflect insincerity or exaggeration, its origins are in the work of Isocrates; and its elaboration during the Renaissance by Erasmus and others (Shrag, 1992, p. 271) gave it its distinctive meaning of argument and persuasion. As Shrag puts it in his discussion of the traditions of knowledge (p. 272):

> The rhetorical tradition realizes the limitations of philosophical argument as a vehicle for persuasion, especially when addressed to those who lack the training to follow the arcane, arid argumentation relished by that tradition. The rhetorical tradition recognizes a fundamental fact, namely, that people are creatures of flesh and blood, of passionate desire and aversion.

I have already made the point that research is about persuasion, and so is rhetoric. My concern is that we come clean about this and recognize rhetoric as part of our professional work. As Shrag notes, rhetoric is a tradition of knowledge that has been "the most influential tradition in European and American schools since the Renaissance" (p. 275). Once we have accepted that research is about persuasion, our task as researchers and graduate supervisors becomes one of acknowledging the place of rhetoric in discussions of the rigor of research, because our students need to know what is rhetoric and what is not, and they need to know what is poor rhetoric and what is good.

We need to replace talk about reliability and validity with a concept that acknowledges that the value and purpose of research lies in human affairs. "Rigor" seems to do this. By promoting the idea of rigor and its facets, we might discourage students of research from reliance upon checklists about reliability and validity. Part of the danger of checklists is that they tend to sanitize research. In contrast, the facets of rigor tell us plainly that research is very human.

I am grateful to Christine Chapman and Jennifer Taylor for help with aspects of this paper.

REFERENCES

American Psychological Association (APA). (1954). Technical recommendations for psychological tests and diagnostics techniques. Psychological Bulletin, 51, 2–7.

Auden, W. H. (1979). Under which lyre. In E. Mendelson (Ed.), W. H. Auden: Selected poems (pp. 178–183). New York: Vintage Books.

Aydelotte, W. (1971). Quantification in history. Reading, MA: Addison-Wesley.

Bruner, J. (1996). Narratives of science. In J. Bruner (Ed.), The culture of education (pp. 115–129). Cambridge, MA: Harvard University Press.

Conant, J. (1957). Harvard case histories in experimental science. Cambridge, MA: Harvard University Press.

Dray, W. (1957). Laws and explanation in history. Oxford, UK: Oxford University Press.

Fenstermacher, G. (1990). Some moral considerations on teaching as a profession. In J. Goodlad, R. Soder, & K. Sorotnik (Eds.), The moral dimensions of teaching (pp. 130–151). San Francisco, CA: Jossey-Bass.

Gergen, M., & Gergen, K. (2000). Qualitative inquiry: Tensions and transformations. In N. Denzin & Y. Lincoln (Eds.), Handbook of qualitative research (2nd ed., pp. 1025–1046). Thousand Oaks, CA: Sage.

Goetz, J., & LeCompte, D. (1984). Ethnography and qualitative design in educational research. New York: Academic Press.

Gould, J., & Kolb, W. (1964). A dictionary of the social sciences. Glencoe, IL: Free Press.

Hempel, C. (1968). Explanation in science and history. In P. Nidditch (Ed.), The philosophy of science (pp. 54–79). Oxford, UK: Oxford University Press.

Heshusius, L. (1994). Freeing ourselves from objectivity: Managing subjectivity or turning toward a participatory mode of consciousness. Educational Researcher, 23(3), 15–22.

Hoepfl, M. (2000). Choosing qualitative research: A primer for technology education researchers. Retreived on Feb. 19, 2001, from <http://www.curriculum.edu.au/tech/articles/choose.htm>.

LeCompte, M., & Schensul, J. (1999). Analyzing and interpreting ethnographic data. Walnut Creek, CA: Altamira Press.

McMillan, J., & Schumacher, S. (1997). Research in education: A conceptual introduction (4th ed.). New York: Longman.

McMillan, J., & Schumacher, S. (2001). Research in education: A conceptual introduction (5th ed.). New York: Addison Wesley Longman.

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

Mischler, E. (1990). Validation in inquiry-guided research: The role of exemplars in narrative studies. Harvard Educational Review, 60, 415–442.

Moschkovich, J., & Brenner, M. (2000). Integrating a naturalistic paradigm into research on mathematics and science cognition and learning. In A. Kelly & R. Lesh (Eds.), Handbook of research design in mathematics and science education (pp. 457–486). Mahwah, NJ: Erlbaum.

Munby, H. (1983). A perspective for analyzing the significance of qualitative research: A response to Richard Heyman. Curriculum Inquiry, 13, 423–427.

Munby, H., Chin, P., & Hutchinson, N. L. (in press). Co-operative education, the curriculum, and "working knowledge." In W. Pinar, W. Doll, D. Truett, & H. Wang (Eds.), The internationalization of curriculum. New York: Peter Lang.

Munby, H., Russell, T., & Martin, A. (2001). Teachers' knowledge and how it develops. In V. Richardson (Ed.), Handbook of research on teaching (4th ed., pp. 877–904): Washington, DC: American Educational Research Association.

Roberts, D. (1982). The place of qualitative research in science education. Journal of Research in Science Teaching, 19, 277–292.

Sandelowski, M. (1993). Rigor or rigor mortis: The problem of rigor in qualitative research revisited. Advances in Nursing Science, 16(2), 1–8.

Shrag, F. (1992). Conceptions of knowledge. In P. Jackson (Ed.), Handbook of research on curriculum (pp. 268–301). New York: Macmillan.

Simpson, J. A., & Weiner, E. S. C. (Eds.). (1989). The Oxford English dictionary (2nd ed., Vols. 1–20). Oxford, UK: Clarendon Press.

Soder, R. (1990). The rhetoric of teacher professionalism. In J. Goodlad, R. Soder, & K. Sorotnik (Eds.), The moral dimensions of teaching (pp. 35–86). San Francisco, CA: Jossey-Bass.

Webb, R. B., & Glesne, C. (1992). Teaching qualitative research. In M. LeCompte, W. Millroy, & J. Preissle (Eds.), The handbook of qualitative research in education (pp. 775–776). New York: Academic Press.

HUGH MUNBY
Faculty of Education
Queen's University, Kingston
Ontario, Canada K7L 3N6