THE TEACHING OF MATHEMATICS
2005, Vol. VIII, 2, pp. 73–81

TOWARDS BASIC STANDARDS FOR RESEARCH IN MATHEMATICS EDUCATION

Djordje Kadijevich

Abstract. Although there seems to be a consensus among mathematics educators that judgments of the quality of research should be based on a set of widely accepted criteria applied in a constructive, non-dogmatic way, such criteria are not yet available. Instead, because a common vocabulary has not been used, various sets of criteria for research quality have been proposed, and no one has attempted to uncover those latent issues concerning quality whose realization and a step-by-step utilization would be particularly beneficial for novice researchers. This paper proposes the grouping of criteria for the quality of research in mathematics education into three basic comprehensive standards. The appropriateness of the three standards is established through an analysis of a representative sample of proposed sets of criteria. The standards are examined in terms of suitable indicators and then applied to the realization of a study searching for the dimension of mathematics attitude that mainly influences mathematics achievement. Although the standards may appear to apply only to the evaluation of research reports, their use may also increase the quality of the design and management of research activities for both quantitative and qualitative research studies.

ZDM Subject Classification: B10; *AMS Subject Classification*: 00 A 35

Key words and phrases: Standards, research, mathematics education.

Introduction

Discussions of the quality of research in education—and mathematics education, in particular—have been going on for more than three decades (e.g., Coburn [4]; Hanna [8]; Howe & Eisenhart [10]; Kilpatrick [12]; Sierpinska [21]; Lester & Lambdin [14]; Romberg [17]; Schoenfeld [18]; Simon [22]; Thompson [23]). These discussions have recognized and respected, among other things, the increasing application of non-statistical methods. In education, where statistical and non-statistical methods are today widely applied within various traditions, high-quality research studies are still rather rare (Burkhardt & Schoenfeld [3]).

Judgments of the quality of research ought to be based upon acknowledged criteria (see, e.g., Lester & Lambdin [14]), but in the language used to describe the various sets of criteria that have been proposed, there seems to be no commonly accepted vocabulary. Also, there are issues associated with research quality that new or relatively inexperienced researchers should be aware of. This study examines whether the proposed criteria for research quality might be incorporated into a small

An extended version of a contribution to 11th Congress of Mathematicians of Serbia and Montenegro, Petrovac na moru, September 28–October 3, 2004.

number of standards, what the indicators of those standards might be, and how they might be applied.

Proposed criteria for research quality grouped under three standards

This study examined the sets of criteria for judging the quality of research in mathematics education proposed by the following authors: Thompson [23], Kilpatrick [12], Sierpinska [21], Dörfler [6], Lester and Lambdin [14], Hart [9], Hanna [8], Schoenfeld [18], and Simon [22]. These criteria could be grouped under three standards: (a) *research relevance*, relating to the extent to which the research being examined is relevant (related or applicable) to mathematics education; (b) *research significance*, denoting the extent to which the research advances the knowledge of the field; and (c) *research rigor*, referring to the extent to which the empirical or theoretical analysis concerning research questions and goals is rigorously and precisely designed, realized, and grounded in evidence based on external data (from other studies) and internal data (from the study itself). Although one might be tempted to equate research significance with overall research quality, the two concepts are different. For example, the quality of a study might be high, but its significance might nonetheless be low, or vice versa. Table 1 shows how the criteria proposed by various authors were classified under the three standards. Note that some of the proposed criteria refer to issues that are covered by more than one basic standard.

Table 1. Criteria for research quality classified by standard Author Target of criteria

Author	**Target of criteria**	**Relevance**	**Significance**	**Rigor**	**Note**
Dörfler (1993)	Research report	Concerned with teaching/learning of mathematics	Novelty, embeddedness in literaure and existing reasearch	Explicate *problématique*, paradigm, and method; justify statements; separate results from interpolation; clarify concepts	Labels for criteria are not given
Hanna (1998)	Research report	Usefulness	Originality	Presence of research paradigm	Criterion of readability and length is not classified
Hart (1998)	Various types of research studies	There is a theory	There is a problem and a theory	There is a problem; there is evidence or data; the work can be replicated; there is a theory	Criterion that the work is reported is not classified

Kilpatrick (1993); Sierpinska (1993)	Any kind of research	Relevance, relatedness	Originality, predictability, objectivity; also relevance	Validity, objectivity, rigor and precision, reproductibility	
Lester and Lambdin (1998)	All types of research acitivties	Wothwhileness	Wothwhileness	Coherence, competence, openness, credibility	Criterion of ethic is not classified
Schoenfeld (2000)	Any empirical or theoretical work	Descriptive power, explanatory power, predictive power, scope	Descriptive power, explanatory power, scope, predicitive power	Rigor and speicifity falsifiability, replicability, multiple sources of evidence	
Simon (2004)	Any empirical or theoretical work	Search for a new construct or description of reality	Description is generally not enough, make a contribution to the knowledge of the field, build on the work of others, research questions usually evolve	Empirical research as an argument, theoretical research as an argument, rigorous application of methodologies is not sufficient, research questions usually evolve	Most labels of research quality issues are given
Thompson (1982)	Constructivist clinical research report	Developing a model that explain students' learning of mathematics	Model viability, model sufficiency	Specify a framework for constructing models, make prototypes clear, ground the framework in data, use viable models, consider model sufficiency	Instead of labels, questions for judging quality are given

The standards bear some resemblance to those proposed by Schoenfeld [18], who argued that the impact of research in education should be examined along (at least) the following three dimensions: trustworthiness, generality, and importance. His dimension of importance appears to be related to research significance, his dimensions of trustworthiness and generality seem to capture basic aspects of research rigor, but he has no counterpart to the research relevance standard.

Indicators of the Three Standards

Research Relevance

Because research relevance deals with the extent to which the research being examined is related or applicable to mathematics education, its indicators may be research relatedness (to use Kilpatrick's [12], and Sierpinska's [21], terminology) and research usefulness (to use Hanna's [8], terminology). Whereas research relatedness denotes the strength of the relation of the research in question to mathematics and its teaching and learning, research usefulness is concerned with the possibility of (more or less straightforwardly) applying the research outcome in the classroom. High research relatedness, therefore, does not necessarily imply high research usefulness. And the relevance of the research is usually not explicitly discussed in a research report.

Research Significance

The standard of research significance may be described by three indicators: gap bridging, research embedding, and research novelty. Each of these indicators depends on the other two.

Gap bridging. Research should always introduce, examine, or operationalize issues that have a previously established niche to be occupied (see Miller & Parker [15]). Any research study begins by identifying a gap. It surveys the research context, arguing that the current situation is inappropriate, that previous findings are contradictory, or that certain issues have been unrecognized or neglected. The argument is usually further clarified by means of research questions to be investigated or goals to be achieved. The research then ends by bridging the gap: suggesting answers to the research questions or ways to attain the goals. A study aiming to be significant must thus indicate a gap and provide a solution whereby that gap is, could, or should be bridged. Although this bridging is typically done in some fashion, to adequately realize the gap and its consequent bridging and to evaluate them properly, relevant research issues need to be made fully explicit, which is seldom the case. Such explicitness would be especially beneficial to those undertaking research that lacks or is developing a tradition of quality.

Research embedding. Standing on the shoulders of giants requires that one comprehensively and critically evaluate previous research efforts and relate one's work to them. Because of the Internet and such resources as the Mathematics Didactics (MathDi) database (see `http://www.emis.de/MATH/DI.html`), it is much easier to locate related mathematics education research today than it was some 10 or 20 years ago. Researchers frequently fail to take advantage of such resources, however, tending to operate in their own narrow research circles and ignoring the work of others in the field (Kilpatrick [11]). One reason can be found in the field's high tolerance toward discursive diversity that has usually been misinterpreted as a license for ignoring other researchers' work (Sfard [20]). It should be noted that, because of low research comparability, great redundancy might occur. Guy Brousseau once estimated that about 80% of the research in mathematics education "is reorganizing, reformulating, and problematizing work that has already been

done" (Gjone [7, p. 51]). Although the fraction may not be that large, the extent to which one's research is embedded in the field remains a crucial indicator of its significance.

Research novelty. Much research has been done to expand and reconceptualize existing knowledge. The crucial question in the evaluation of a research study is the following: What has been (or would likely be) found that is new, and how does (or might) this finding advance the field? The author of any research report should have clearly underlined what about the research was novel, but that is frequently not done. Consequently, the research is of limited value, and the report has little chance of appearing in an outstanding journal or other such publication. Because the reader may find this indicator too broad, it should be emphasized that, as Kilpatrick [12] notes in his discussion of research originality, novelty does not rule out a replication or a meta-study. Both types of study can and should extend previous work in some way, especially if the related literature has not provided a definitive answer. Note also that research novelty should not be equated with research originality (as evidenced by conducting a study that has never been done before or by approaching old data or questions in a novel manner; see Lester & Lambdin [14]). The research questions should be original (at least some of them). And the research method and its results may be original. Nevertheless, the main question is whether these findings are new (at least to some extent) and whether they advance the field.

Research Rigor

The standard of rigor is concerned with the exactness of the main steps in the research, from generating questions and goals to answering the questions and attaining the goals. Rigor requires precision of measurement, meaning, and reasoning within a relevant context that makes use of an appropriate method (see Kilpatrick [12]). Any piece of research comprises the following components: a research context, a concept or construct[1] being proposed or used, a research question or goal, a method, the collection and analysis of data (of some kind), and a research outcome (whether an empirical or a theoretical finding, a method, a theory, or an instructional design). For research to be rigorous, each component must be justified. Consequently, the standard of rigor may be described by the following indicators: (a) research questions and aims justification, (b) research concepts and constructs justification, (c) research method justification, (d) data collection and analysis justification, and (e) research outcome justification. The justification of research context should not be a separate indicator of research rigor inasmuch as it is covered by research embedding.

With regard to justification in general, to paraphrase Lehrer [13], Object O is fully justified if and only if it coheres with a system of relevant, already justified objects. In justifying research concepts and constructs, operative definitions are, as Sfard [20] emphasizes, crucial in any kind of research.

[1] A construct should be viewed as a postulated attribute of the learning or teaching of mathematics that is assumed to be reflected in the score of an instrument measuring it (Cronbach & Meehl [5, p. 283]).

Each of the five indicators above should, whenever appropriate, make use of reliability and validity, which are, according to a recent article on qualitative inquiry (Morse, Barrett, Mayan, Olson, & Spiers [16]), still central issues for the rigor of any kind of research. These two notions—adapted to the type of research, of course—are usually defined roughly as follows:

- Reliability refers to the accuracy of the data produced by the application of a particular method (or the extent to which that method would yield the same results when applied to the same objects). It is typically estimated by Cronbach's alpha (for quantitative variables) or Cohen's kappa, the measure of so-called inter-observer agreement (for categorical data used in qualitative studies). For theoretical data, reliability may be reflected in a sort of "inter-concluder" agreement (the degree of agreement between conclusions made by independent researchers), which seems missing in the literature.
- Validity is concerned with the degree to which what is measured is that intended rather than something else. Validity is, according to Anfara, Brown, and Mangione [2], also concerned with the degree to which the conclusions are trustworthy, drawn from the collected data, and matched with reality (internal validity) as well as the degree to which the obtained findings are generalizable to a larger population (external validity). Note that in quantitative research reliability and validity are included in construct validation.[2]

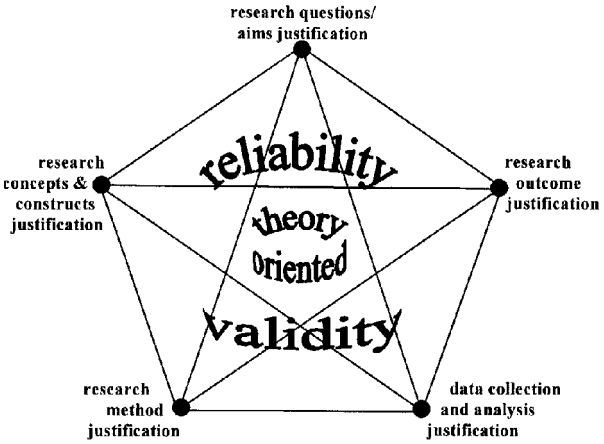

Fig. 1. Framework for the indicators of research rigor.

In consideration of the proposed criteria of coherence, competence, openness, and credibility (Lester & Lambdin [14], each of the indicators should be examined in

[2]Construct validation (or construct justification, which seems a more appropriate term) requires some justification of the content validity, face validity, unidimensionality (homogeneity), reliability, convergent validity, discriminant validity, nomological validity, concurrent validity, and predictive validity of the examined construct (see Ahire & Devaraj [1]).

terms of the others. This approach, when research is planned and conducted, would, as Simon [22] suggests and Morse et al. [16] underline, enable research components to be incrementally refined towards a higher level of research rigor. Furthermore, the arguments of Hart [9] and Morse et al. [16] imply that the research components of most studies should mirror a theory being used or should support the elaboration or development of a theory. Consequently, the indicators of research rigor should be utilized in a framework presented in Figure 1, where, of course, the relations among these indicators are not of the same strength. Note that almost all the proposed criteria classified in Table 1 under the standard of research rigor can be found in such a framework. For example, although Schoenfeld's [18] criterion of replicability seems basically (though implicitly) concerned with some sort of reliability and validity, his criterion of triangulation (multiple sources of evidence) is a key issue of data collection and analysis justification.

Applying the three standards

Applying an unordered set of criteria for research quality, usually in a simultaneous fashion, is a difficult task, especially for novices. The three basic standards should be applied in the order they were introduced above: (a) relevance, (b) significance, and (c) rigor (possibly through a three-tier evaluation, with one tier per standard). There are three reasons for such an approach. First, the examination of relevance is less demanding than that of significance, which is in turn less demanding than that of rigor. (Because of its greater complexity, the examination of research rigor may itself require several tiers of evaluation.) Second, the initial preparation of a research study deals explicitly or implicitly with these standards in the same order, which may, during a later stage of research preparation or of research utilization be extended to a back-and-forth movement between research significance and research rigor (as gap bridging requires stronger justification of research questions, aims, and outcomes) or to an interplay between several elements of research rigor. Third, if one standard in the sequence is poorly met, there may be no need to consider the applicability of the next one (until matters improve). Of course, the quality of some published research is usually fully visible only after several years when, among other issues, citations of that research can be examined.

To illustrate the application of the three standards briefly, let us suppose that a study searching for the dimension of mathematics attitude that mainly influences mathematics achievement is to be realized.

Research relevance. Although this study is highly relevant to teaching of mathematics (knowing which dimension, if any, of mathematics attitude primarily influences mathematics achievement does have important educational consequences), the criterion of research usefulness is not met because the study does not deal with a successful approach strengthening the uncovered dimension, which, provided that such a dimension exists, is an important direction for further research.

Research significance. This standard may apply to a considerable extent. The gap is clearly formulated and would eventually (through strengthening research rigor) be bridged in a skilled way. Let us suppose that the study is fully embedded in

the field (that all relevant references are utilized) and that the dimension in question has not been uncovered (that the references up do date have only suggested what the dimension is likely to be, for example). There is no doubt that the outcome of this study, especially if that dimension is found, would bring to light a new piece of knowledge, which may advance the field though, for example, developing an approach that can strengthen the uncovered dimension.

Research rigor. To attain this standard at a level that is usually expected, this study needs, among other things, to select relevant dimensions of mathematics attitude, to operationalize these dimensions by an appropriate instrument, to establish its factor validity, to confirm the reliability of the measure of each dimension, to apply appropriate statistical analyses (in finding relevant correlation coefficients and comparing them), and to justify the outcome by relating it to relevant previous research in the field. The more justified each of these requirements is, the more research rigor is attained. To attain the standard of research rigor at an advanced level, large representative samples from different countries are to be analyzed, for example, hoping that a general pattern would emerge. Because using samples from different countries may generate various problems (e.g. factor validity not applicable for each country and/or low reliabilities of the measures of the examined dimensions for some dimensions and some countries), there may be a need to apply a suitable data transformation that eliminates or diminishes these problems.

Coda

Although the ethics of research is not covered by the proposed standards, they do support the main aspects of the evaluation of research quality, which can be done incrementally before, during, and after the research is conducted. If an evaluation is undertaken while the research is planned and conducted, each of the indicators, when applied, should be reflected in an appropriate activity, which ought to increase research relevance, significance, or rigor. Although an awareness of being evaluated according to well-defined standards might push some researchers to perform rituals rather than engage in genuine inquiry, the development of basic research standards, a clarification of their indicators and how those indicators are to be applied, and the widespread use of these standards would undoubtedly strengthen and unify efforts to improve our research and advance our field. The author hopes that this contribution, which should primarily be viewed as an introductory account on this important topic, would eventually help us develop and utilize widely-agreed basic research standards.

REFERENCES

1. Ahire, S. L., & Devaraj, S., *An empirical comparison of statistical construct validation approaches*, IEEE Transactions on Engineering Management, **48** (2001), 319–329.
2. Anfara, V. A., Brown, K. M., & Mangione, T. L., *Qualitative analysis on stage: Making the research process more public*, Educational Researcher, **31(7)** (2002), 28–38.
3. Burkhardt, H., & Schoenfeld, A. H., *Improving education research: Toward a more useful, more influential, and better-funded enterprise*, Educational Researcher, **32(9)** (2003), 15–21.

4. Coburn, T. G., *Criteria for judging research reports and proposals*, Journal for Research in Mathematics Education, **9** (1978), 75–78.
5. Cronbach, L. J., & Meehl, P. E., *Construct validity in psychological tests*, Psychological Bulletin, **52** (1955), 281–302.
6. Dörfler, W., *Quality criteria for journals in the field of didactics of mathematics*, In: G. Nissen & M. Blomhøj (Eds.), Criteria for scientific quality and relevance in the didactics of mathematics (pp. 75–87), Roskilde, Denmark: Danish Research Council for the Humanities, 1993.
7. Gjone, G., Review of the book *Theory of didactical situations in mathematics, Didactique des mathematiques, 1970–1990*, Nordic Studies in Mathematics Education, **7** (1999), 47–52.
8. Hanna, G., *Evaluating research papers in mathematics education*, In: A. Sierpinska & J. Kilpatrick (Eds.), Mathematics education as a research domain: A search for identity, New ICMI Study Series, **4**, 399–407, Dordrecht, Kluwer, 1998.
9. Hart, K. M., *Basic criteria for research in mathematics education*, In: A. Sierpinska & J. Kilpatrick (Eds.), Mathematics education as a research domain: A search for identity, New ICMI Study Series, **4**, 409–413, Dordrecht, Kluwer, 1998.
10. Howe, K. R., & Eisenhart, M. A., *Standards for qualitative (and quantitative) research: A prolegomenon*, Educational Researcher, **19(4)** (1990) 2–9.
11. Kilpatrick, J., *Editorial* Journal for Research in Mathematics Education, **18** (1987), 82.
12. Kilpatrick, J., *Beyond face value: Assessing research in mathematics education*, In: G. Nissen & M. Blomhøj (Eds.), Criteria for scientific quality and relevance in the didactics of mathematics (pp. 15–34), Roskilde, Danish Research Council for the Humanities, 1993.
13. Lehrer, K., *Theory of Knowledge*, London: Routledge, 1992.
14. Lester, F. K., & Lambdin, D. V., *The ship of Theseus and other metaphors for thinking about what we value in mathematics education research*, In: A. Sierpinska & J. Kilpatrick (Eds.), Mathematics education as a research domain: A search for identity, New ICMI Study Series, **4**, 415–425, Dordrecht, Kluwer, 1998.
15. Miller, T., & Parker, D., *Writing for the reader: A problem solution approach*, English Teaching Forum, **35(1)** (1997), 2–7. Retrieved September 1, 2005, from http://exchanges.state.gov/forum/vols/vol35/no1/p2.htm.
16. Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J., *Verification strategies for establishing reliability and validity in qualitative research*, International Journal of Qualitative Methods, **1(2)** (2002), Article 2. Retrieved September 1, 2005, from http://www.ualberta.ca/˜ijqm/
17. Romberg, T. A., *Criteria for evaluating educational research*, Investigations in Mathematics Education, **2** (1969), x–xi.
18. Schoenfeld, A., *Purposes and methods of research in mathematics education*, Notices of the American Mathematical Society, **47** (2000), 641–649.
19. Schoenfeld, A. H., *Research methods in (mathematics) education*, In: L. English (Ed.), Handbook of international research in mathematics education (pp. 435–488), Mahwah, NJ, Erlbaum, 2002.
20. Sfard, A., *What could be more practical than good research? On mutual relations between research and practice of mathematics education*, Educational Studies in Mathematics, **58** (2005), 393–413.
21. Sierpinska, A., *Criteria for scientific quality and relevance in the didactics of mathematics*, In G. Nissen & M. Blomhøj (Eds.), Criteria for scientific quality and relevance in the didactics of mathematics (pp. 35–74), Roskilde, Danish Research Council for the Humanities, 1993.
22. Simon, M., *Raising issues of quality in mathematics education research*, Journal for Research in Mathematics Education, **35** (2004), 157–163.
23. Thompson, P. W., *Were lions to speak, we wouldn't understand*, Journal of Mathematical Behavior, **3(2)** (1982), 147–165.

Megatrend University and Mathematical Institute SANU, Belgrade, Serbia
E-mail: djkadijevic@megatrend-edu.net