Scientifically Based Research in Quantitative Literacy: Guidelines for Building a Knowledge Base

Richard L. Scheaffer
University of Florida, rls907@bellsouth.net

Scientifically Based Research in Quantitative Literacy: Guidelines for Building a Knowledge Base

Abstract

Research in quantitative literacy (QL) is in its infancy, so now is the time to begin a regimen for healthy growth into adulthood. As a new discipline still defining itself, QL has the opportunity to build a sound infrastructure for accumulating a solid body of interconnected research that will serve the discipline well in years to come. To that end, much can be learned from recent studies of the weaknesses of mathematics education research and recommendations on how to overcome them. Mathematics education lacks a strong research foundation, one that is scientific, cumulative, interconnected, and intertwined with teaching practice. These weaknesses can be alleviated by following a model built around five key components of a high-quality research program: generating ideas, framing those ideas in a research setting, examining the research questions in small studies, generalizing the results in larger and more refined studies, and extending the results over time and location. Single research projects having only one or two of these components should link to others so that a viable research program that is interconnected and cumulative can be identified and effectively used to improve both teaching practice and future research. Detailed reporting guidelines for each component of the model are outlined in the following sections.

Keywords

mathematics education research, scientifically based research

Creative Commons License

Introduction

> No one would think of getting to the Moon or of wiping out a disease without research. Likewise, one cannot expect reform efforts in education to have significant effects without research-based knowledge to guide them.
>
> > (Shavelson and Towne, *Scientific Research in Education* [National Research Council 2002], 1)

> The central idea of evidence-based education—that education policy and practice ought to be fashioned based on what is known from rigorous research—offers a compelling way to approach reform efforts.
>
> > (Towne, Wise and Winters, *Advancing Scientific Research in Education* [National Research Council 2004], vii)

> The teaching and learning of mathematics in U.S. schools is in urgent need of improvement. The nation needs a mathematically literate citizenry, but most Americans graduate from high school without adequate mathematical competence.
>
> > (Ball, *Mathematical Proficiency for All Students* [RAND Mathematics Study Panel 2003], xi)

It is widely recognized that sound reform of education policy and practice must be based on sound research. For the hotly debated field of mathematics education the reality is, however, that whatever claims of strengths or weaknesses in programs and recommendations for change may be debated, the research that backs up the claims is quite likely to be diffuse and only moderately compelling at best. The fledgling field of quantitative literacy (QL) can and should learn valuable lessons from what has happened in mathematics education as it begins to build a research base and infrastructure to support improved QL education—at both the school and college levels—in the years to come.

Mathematics education researchers are well aware of the lack of an adequate research base in many areas and are working to improve the situation. That spirit led to a series of workshops to investigate how mathematics education researchers and statisticians could strengthen scientifically based research in mathematics education by sharing ideas from their respective disciplines. The result of the workshops is the report *Using Statistics Effectively in Mathematics Education Research* (USEMER) (American Statistical Association 2007).

Funded by the National Science Foundation, workshops were held over a three-year period, each with about twenty participants nearly equally divided between mathematics educators and statisticians. In these exchanges the mathematics educators presented honest assessments of the status of mathematics education research (both its strengths and its weaknesses), and the statisticians

provided insights into modern statistical practices and methods that could be more widely used in such research. The discussions led to an outline of guidelines for evaluating and reporting mathematics education research, which were molded into the report referenced above. The purpose of these guidelines is to foster the development of a stronger foundation of research in mathematics education, one that will be scientific, cumulative, interconnected, and intertwined with teaching practice.

These are worthwhile goals for QL research as well. QL is starting out almost from square one; it is essential that this new educational field build academic credentials that will stand up to the tough scrutiny of education research that is becoming ever more challenging. Following an adaptation of the guidelines proposed for mathematics education research will aid in this process. This paper is based on the USEMER guidelines, with emphasis on and expansion of those points particularly relevant to QL research.

Scientifically Based Research

What is scientific research in education? The No Child Left Behind Act (NCLB) of 2001 has many problems and no shortage of critics, but perhaps one of its positive contributions was the attempt to define *scientifically based research*. NCLB defines scientifically based research as "research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs." Such research

- Employs systematic empirical methods that draw on observations, sample surveys, or experimentation;
- Involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;
- Relies on measurements or observational methods that provide reliable and valid data across evaluators and observers, across multiple measurements and observations, and across studies by the same or different investigators;
- Is evaluated, as appropriate, using qualitative, quantitative, exploratory, experimental, or quasi-experimental designs, with random assignment being preferred for studies that attempt to make generalizations to broad populations;
- Ensures that experimental studies are presented in sufficient detail and clarity to allow for replication of both the experiment and the analyses.

(*No Child Left Behind Act of 2001*, http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf, Section 115 STAT.1964)

This definition, along with the reports quoted above, provided a starting point for building the guidelines of USEMER, which provides practical guidance on how these (largely statistical) requirements of NCLB can be addressed. A summary of the USEMER guidelines, with comments on their relevance to QL research, is provided in the following sections.

Components of a Research Program

The USEMER guidelines are built around a model involving five key components essential to a high-quality research program: *generating* ideas, *framing* those ideas in a research setting, *examining* the research questions in small studies, *generalizing* the results in larger and more refined studies, and *extending* the results over time and location. Any single research project may have only one or two of these components, but such projects should link to others so that a viable research program covering all components can be developed. In turn, such an interconnected and cumulative program can lead to improvement of both teaching practice and future research. In order to provide a useful database and a sound infrastructure for research it is essential that such linkages occur.

Research ideas, often generated from practice or the research of others, typically begin as loosely formed questions or ideas that should be explored in some detail before being framed into researchable hypotheses. Many studies suffer from fuzzy data being collected in a haphazard manner for loosely defined purposes because the "generating" and "framing" steps were not studied conscientiously. A "frame" clarifies the *goals* of the research program and defines the *constructs* it entails, formulates the tools and procedures for the *measurement* of those constructs, and outlines the *logistics* needed to put the ideas into practice and study their *feasibility*. These framing issues should be developed interactively as researchers decide what the program's initial research questions or hypotheses should be and how studies might best be shaped and managed.

A model of how the components are intended to fit together is provided in Figure 1. The five components appear here in linear fashion, but in practice there should be many feedback loops that feed into earlier parts of a program; a good research program is cyclical rather than linear. For example, the generate-frame cycle may be repeated a number of times before a research project goes on to the examine stage.

Once goals are clear, constructs are defined, adequate measures are in hand, and a study looks to be feasible in terms of time, effort, personnel and expected outcomes, a project is ready to be examined in a relatively small-scale investigation, usually within one or two institutions. The main purpose here is to
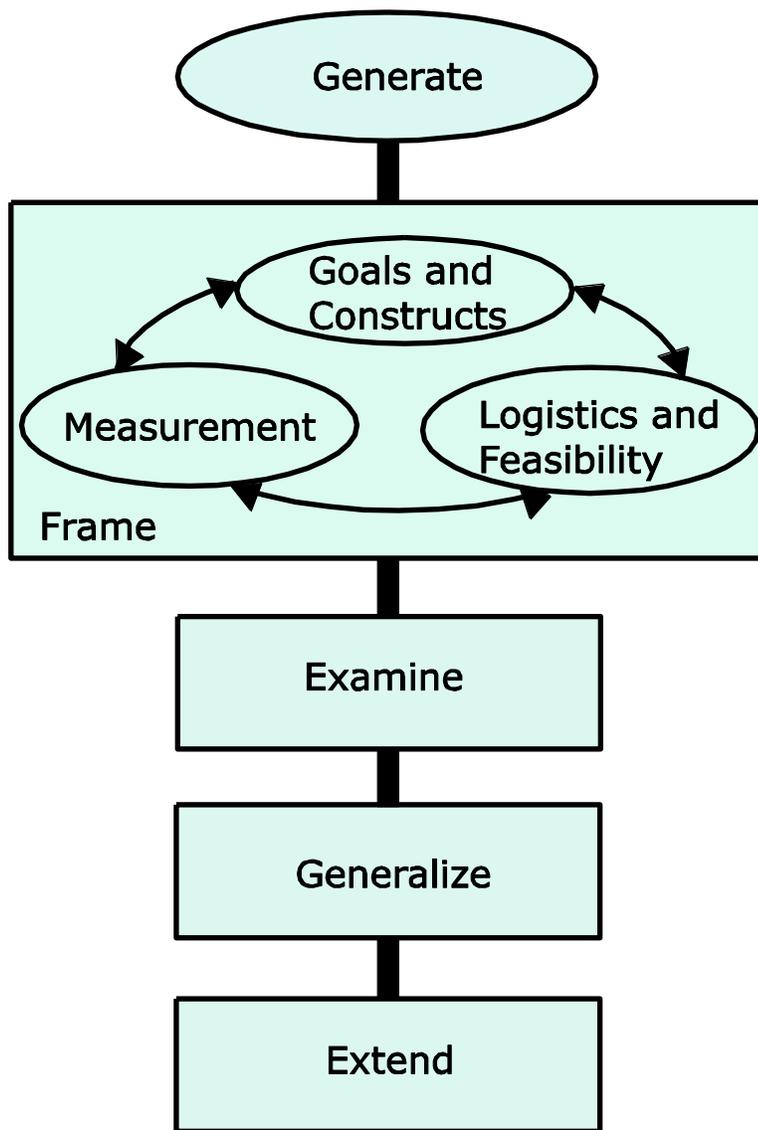
Figure 1. Structure and components of a research program

gain a clearer understanding of the phenomenon under investigation and to see what might work in various settings to generate sound data on the original hypotheses. In the process, this step often cycles back to address one or more framing issues. (Much of the research currently available on issues related to QL is at either the framing or examining levels.)

Successful examination in small-scale studies can then lead to attempts to generalize the results to large-scale studies, usually across multiple institutions

and settings. If the research is an intervention study, then this is the stage to seriously consider using randomized controlled trials to allow cause and effect conclusions across the settings of the experiment. Otherwise, any conclusions drawn can only speak of possible associations.

There is one final step for a program that has successfully completed a generalization. The results may be extended further by considering other settings in time or location for the research, or synthesizing the results with other research projects of a similar nature. At the very least, the results should be fed back to the framing stage of related projects so that investigators can study the implication for future research.

Reporting Guidelines for Research Components

In designing and conducting a research program, or projects within a program, one should have in mind a set of guidelines that should be followed (a checklist, if you will) so that the research has a good chance of producing results that are scientifically defensible, repeatable, able to be communicated to others in an unambiguous manner, and help build the knowledge base in QL research. One good way to focus these guidelines is to think in terms of "What should I report to others about this research to make its intentions and results so clear that others could repeat the research or the analysis?" Such *reporting guidelines* are considered in detail in the USEMER report; what follows is a brief sampling of key portions.

Framing Component 1: Goals and Constructs
Reporting Guidelines
- State the research question and identify and describe the research in related fields.
- Define the variables and measures used.
- Describe the basic research that will guide the research project, showing how the proposed research will fill gaps in the accumulated knowledge.
- Provide exploratory and descriptive statistics with graphical representations, if appropriate, with interpretations to support the background and setting of the proposed research. (Attempts at rigorous statistical inference are neither needed nor appropriate at this stage.)

Framing Component 2: Measurement
Reporting Guidelines
- Provide a summary of the literature review regarding relevant measures.
- Provide key details regarding development of new measures, and/or selection of "off-the-shelf" measures.

- For all measures, report how the variable is operationalized and measured and what relationships the variable has with other variables used in the research.
- For all measures, report evidence of validity, reliability, and fairness that is specifically relevant to the context in which the measure will be used.

Measurement issues will be considered in greater detail in the next section.

Framing Component 3: Logistics and Feasibility
Reporting Guidelines
- Describe the study design of the project.
- Describe the population of interest versus the sample studied, including demographic characteristics.
- Describe the method of sampling (if used).
- Identify the sampling unit and the unit of analysis.
- Describe the treatment (if used) and measures in enough detail to allow replication.
- Report empirical data in a complete fashion, including data on the characteristics of subjects.
- Provide descriptive statistics and graphical representations; rigorous statistical inference is not needed and is most likely unwarranted at this stage.
- Address confidentiality and consent issues.

Examining the Research Program
Reporting Guidelines
- Provide enough information on the study design to allow replication of the study.
- Provide a thorough description of the data analyses so that they could be replicated.
- Report characteristics of measures, including reliability, bias, and validity.
- Summarize the informed consent process, the percent of potential subjects consenting, and any related human subjects ethical issues.

If formal statistical inference is warranted, see the list provided in the USEMER report; the items listed there are an expansion of those listed below for the generalizing component.

Generalizing the Research Program
Reporting Guidelines
- Describe the research program and the materials being tested.
- Summarize the informed consent process, the percent of eligible subjects consenting, and any related human subjects ethical issues.
- List testable research hypotheses and translate them into statistical hypotheses.

- Specify the type of study design that addresses the hypothesis (experiment, quasi-experiment, matching, repeated measures, etc.).
- If sampling is used, define the population of interest, the exclusion/inclusion criteria for obtaining the sample, and the sampling unit.
- Identify the unit of randomization and the unit or units of analysis.
- Describe potential sources of biases and measures taken to minimize bias.
- Address the sample size or power calculation, effect-size specification (reasons for choice of sample size, including reflections on power, error rate control, etc.).
- Describe the statistical methods of analysis employed.
- State assumptions and describe the methods used to check if they hold and to assess sensitivity if they do not hold (under modest perturbations).
- Summarize the results of appropriate tests of assumptions.
- Provide appropriate graphical or tabular representations, including sample sizes and measures of variability.
- Provide appropriate summary statistics and statistical tables with sufficient information to replicate the analysis.
- Provide enough information to allow replication of the study methods and procedures.
- Provide, if allowed, access to unidentified data with appropriate confidentiality safeguards in place.

Extending the Research Program

Reporting Guidelines

- Describe the research program being studied.
- Describe the nature of the long-term study (experimental, quasi-experimental, sample survey, observational).
- Describe the goals, methods and procedures of the study (monitoring for changes in implementation, process improvement, gather information for new intervention study, etc.).
- Describe the data being collected and provide appropriate summaries.
- Provide the statistical inferences with attention to all applicable details from the guidelines for generalizing research.

Measurement

An essential component of a successful education research project or program is to have good measures of the phenomenon being investigated. But what is a "good" measure? Traditionally, measures of educational concepts and constructs (sometimes called assessment variables, as compared to non-assessment variables like age, years of education, ethnicity, and so on) are evaluated on the basis of their validity and reliability, but a related component, fairness, is sometimes considered as a separate category.

- *Validity* is broadly defined as the extent to which a measure is meaningful, relevant, and useful for the research at hand.

- *Reliability* is broadly defined as the extent to which the measure is free of random error.

- *Fairness* is broadly defined as the extent to which the implementation of the measure is free of systematic error that would undermine validity for one or more subgroups.

See the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999) for further details on these concepts. The broad definitions presented above are generalizations of perhaps the more standard definitions of validity, reliability and fairness such as:

> When stakes are high, it is particularly important that the inferences drawn from an assessment be *valid, reliable,* and *fair.* … Validity refers to the degree to which evidence and theory support the interpretations of assessment scores. Reliability denotes the consistency of an assessment's results when the assessment procedure is repeated on a population of individuals or groups. And fairness encompasses a broad range of interconnected issues, including the absence of bias in the assessment tasks, equitable treatment of all examinees in the assessment process, opportunity to learn the material being assessed, and comparable validity (if test scores underestimate or overestimate the competencies of members of a particular group, the assessment is considered unfair).
>
> (Pellegrino, Chudowsky and Glaser, *Knowing What Students Know* [National Research Council 2001], 39)

Most variables of interest can be operationalized in many different ways; evidence of learning in mathematics or QL could be measured by a quiz score, a judgment score based on an interview by a teacher, or a rubric that takes into account various stages of working out a solution to a problem. Thus, validity, reliability and fairness must be considered in the context of the operational definition and its use, and are not inherent properties of a variable itself.

The following, taken directly from USEMER, summarizes the development and reporting requirements of measures.

> For every assessment that is used in every research process, it is essential to develop and report as appropriate:

1. information on the construct or behavioral domain that the assessment is intended to measure in the specific research process, the alignment of that construct with the goals of the research, and the limitations of the assessment in this context;
2. information about the sample or population to which the assessment will be administered, the circumstances of administration or implementation of the assessment (e.g., physical setting, time limits), and ways in which these are similar to or different from the setting in which published validity, reliability, and fairness evidence (if any) was obtained; and
3. evidence of validity, reliability, and fairness that is specific to the setting in which the assessment is administered, the particular population to which it is administered, the way it is scored, and the use to which the scores are put.

Establishing or verifying this information is no small task, especially if the measures are being developed from scratch. For this reason and for the purpose of comparing results across studies, it is preferable to use "off the shelf" measures that have a track record of good validity and reliability that can be referenced over measures that are "home grown," even if the developed instruments need some minor modification in order to fit the study at hand. For example, the Balanced Assessment in Mathematics program, a joint effort of Michigan State University, UC Berkeley, Harvard and the Shell Centre in the UK, provides well-tested measures in mathematical reasoning and skills.[1] The Assessment Resource Tools for Improving Statistical Literacy (ARTIST) project is building tested measures in statistical reasoning and skills.[2]

Observational versus Experimental Research

Consider the difference between the following two research scenarios. Investigator 1 collects quiz scores on a class of 6th grade students who have gone through Series *A* of lessons on percent and then does the same for another 6th grade class that has gone through Series *B* of lessons on percent with the same teacher. The classes are made up of those students assigned to the specific periods at the start of the school year. In another school Investigator 2 gets permission to take two 6th grade classes with flexible schedules and randomly assign half to Series *A* and half to Series *B*, with the same teacher. This seemingly small difference in design makes the first investigation an observational study and the second a randomized experiment. If *A* students score better than *B* students in the observational study of Investigator 1 all that can be concluded is that there may be an association between performance and the lesson

---

[1] "Balanced Assessment" http://balancedassessment.concord.org.
[2] ARTIST Website, https://app.gen.umn.edu/artist/

series, but the observed difference may be due to many other possible factors such as student ability, motivation and so on. If *A* students score better than *B* students in the designed experiment of Investigator 2 it can be concluded (with appropriate measures of error) that there is evidence of Series *A* causing an improved performance. The causal statement is allowed because the uncontrolled factors like ability and motivation are somewhat balanced for the two groups by the randomization process (especially so if the number of students per group is fairly large).

Of course, known relevant factors in a study should be controlled by more elaborate experimental designs. For example, if two teachers are to be used, then each teacher should teach a group of students with each series so that the lesson series effects can be separated from teacher effects. If all Series A students are taught by one teacher and all Series B students by another, it is impossible to determine if observed differences in student performance are due to the teacher or the lesson series. In any case, the important point is that appropriate randomization allows one to make causal statements while observational studies generally allow only statements of possible association. But, randomized studies tend to be larger, more complicated to conduct, and more expensive than non-randomized studies. Thus, they should be used only when the framing and examining components show that a study portends real promise with adequately tested measures.

It should be duly noted that the unit of randomization need not be an individual student. Sometimes it is more appropriate to randomize on classes, schools or other groups of subjects (majors, perhaps). These group randomization plans will inevitably require more subjects because the power of a statistical analysis to detect significant differences or trends depends on the number of units randomized, not the total number of subjects in the study.[3]

Building a Knowledge Base for QL Research and Teaching

As the title suggests, the purpose of this paper is to provide a model, with guidelines on how to use it effectively, for research programs that will have the capability of contributing to a comprehensive knowledge base in QL. The author recommends the following as key steps toward building such a knowledge base.

---

[3] In somewhat technical parlance, the number of units randomized determines the degrees of freedom for a statistical inference procedure. See Raudenbush (2005) for more on methodology in education research.

1. Cooperation and communication among researchers

An essential ingredient to building a strong base of knowledge in QL and an infrastructure that supports and sustains it is to think in terms of cooperative ventures and open communication among all parties involved in the QL research enterprise, including those who may make use of the research in their teaching. The QL research community should set up an efficient and effective process for keeping individual researchers and research teams informed on current research. Such information could be used to foster larger collaborative efforts among research teams, much as is done in medical research and in areas of the social sciences.[4] A strong QL Collaboration could then become the central agency to archive and disseminate information about QL research and how that research relates to practice.[5]

2. Categorization of research papers according to the proposed model

At least for papers published in *Numeracy* and more broadly if possible, authors and editors should agree on which of the five components of the model a published paper best matches. Such a categorization scheme would provide an easy way for the QL community to see, for example, if any progress is being made toward generalized research on "best practices" in a quantitative reasoning course for college students, or what papers are available on developing measures for assessing student learning in QL. This could also expedite the formation of teams of researchers to move individual research projects into a comprehensive research program.

3. Reporting guidelines

The reporting guidelines outlined in this paper and presented in more detail in the USEMER report should be taken seriously by authors, reviewers and editors so that some consistency in both quality and style of QL publications can be achieved. Such consistency would expedite steps 1 and 2 above.

4. Data availability

Because data sharing is basic to the notion of accumulating a body of knowledge, every research paper in QL should make available the data on which the research is based so that others can reanalyze it and build upon it. In some disciplines (economics, for example), certain journals require that accepted papers be accompanied by the data either for print or posting on a web site. In QL research, this practice could be facilitated through the *National Numeracy Network*. Being

---

[4] Campbell Collaboration Web site for the social sciences: http://www.campbellcollaboration.org/
[5] See Burkhardt & Schoenfeld (2003) for more on building infrastructure in education research.

an electronic journal, *Numeracy* has the capability of managing data files in a user-friendly manner; it should take advantage of this by requiring that data files be provided in the appendices of papers. This could well be the first step toward establishing a QL Collaboration within the *National Numeracy Network*, a Collaboration that fosters communication, teamwork, relevance and open access for all who have an interest in quantitative literacy.

Conclusion

 "The challenge faced by school mathematics in the United States today—to achieve both mathematical proficiency and equity in the attainment of that proficiency—demands the development of new knowledge and practices that are rooted in systematic, coordinated, and cumulative research" (RAND Mathematics Study Pattern 2003, 5). To keep from falling into the same patterns that have prevented mathematics education research from claiming the heights that it should, QL research must strive for a strong base of systematic, coordinated and cumulative research from the outset. Appropriate guidelines for reporting and evaluating such research need to be adopted by the community of researchers so that such a stable base can be achieved. If such achievement is accomplished, research in QL education may avoid qualifying for criticisms such as the following by Arthur Levine, current President of the Woodrow Wilson Foundation.

> As a nation, the price we pay for inadequately prepared researchers and inadequate research is an endless carousel of untested and unproven reform efforts, dominated by the fad du jour. Ideology trumps evidence in formulating educational policy. And our children are denied the quality of education they need and deserve. (Levine 2007, 71)

Acknowledgments

References

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

American Statistical Association. 2007. *Using Statistics Effectively in Mathematics Education Research*. Working Group on Statistics in Mathematics Education Research, R. Scheaffer, Chair. http://www.amstat .org/education/pdfs/UsingStatisticsEffectivelyinMathEdResearch.pdf .

Burkhardt, H., & Schoenfeld, A. H. 2003. Improving educational research: toward a more useful, more influential, and better funded enterprise. *Educational Researcher* 32(9), 3-14. http://dx.doi.org/10.3102/0013189X 032009003

Feuer, M., P. W. Holland, B. F. Green, M. W. Bertenthal and F. C. Hemphill, eds. 1999. *Uncommon measures: Equivalences and linkage amongst educational tests*. National Research Council, Commission on Behavioral and Social Sciences and Education, Board on Testing and Assessment, Committee on Equivalency and Linkages of Educational Tests. Washington, DC: National Academy Press. http://www.nap.edu/catalog.php?record_id =6332

Levine, Arthur. 2007. *Educating researchers*. Washington DC: The Education Schools Project. http://www.edschools.org/EducatingResearchers/ educating_researchers.pdf

Pellegrino, J., N. Chudowsky and R. Glaser, eds. 2001. *Knowing what students know*. National Research Council, Center for Education, Board on Testing and Assessment, Committee on the Foundations of Assessment. Washington, DC: National Academy Press. http://www.nap.edu/catalog.php?record_id =10019

RAND Mathematics Study Panel. 2003. *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education*. RAND Mathematics Study Panel , Deborah L. Ball, Chair. U.S. Department of Education, Office of Educational Research and Improvement, MR-1643.0-OERI. Santa Monica, CA: RAND. http://www.rand.org/pubs/monograph_reports/MR1643/

Raudenbush, S. W. 2005. Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher,* 34(5), 25–31. http://dx.doi.org/10.3102/0013189X034005025

Shavelson, R.J., and L. Towne, eds. 2002. *Scientific research in education*. National Research Council, Center for Education, Division of Behavioral and

Social Sciences and Education. Committee of Scientific Principles of Education Research. Washington, DC: National Academy Press. http://www.nap.edu/catalog.php?record_id=10236

Towne, L., L. L. Wise, and T. M. Winters, eds. 2004. *Advancing scientific research in education*. National Research Council, Center for Education, Division of Behavioral and Social Sciences and Education. Committee on Research in Edcuation. Washington, DC: National Academy Press. http://www.nap.edu/catalog.php?record_id=11112