7. Report of the Subcommittee on Standards of Evidence

Valerie F. Reyna, Chair
Camilla Persson Benbow
A. Wade Boykin
Grover “Russ” Whitehurst

DRAFT 3/6/08

CONTENTS

I. Background: Categories of Internal and External Validity
II. Quantity, Quality, and Balance of Evidence
  A. Strong Evidence
  B. Moderately Strong Evidence
  C. Suggestive Evidence
  D. Inconsistent Evidence
  E. Weak Evidence
III. Applying the Criteria
IV. Task Group Guidelines
  A. Learning Processes Task Group
  B. Conceptual Knowledge and Skills Task Group
  C. Instructional Practices Task Group
  D. Teachers Task Group
V. Procedures
  A. Screening Criteria for the Literature Search
  B. Documenting the Quality of the Evidence Used in the Report
VI. Recommendations

BIBLIOGRAPHY

NATIONAL MATHEMATICS ADISORY PANEL: GUIDELINES FOR STANDARDS OF EVIDENCE

The President's Executive Order calls for the National Mathematics Advisory Panel to marshal the best available scientific evidence and offer advice on the effective use of the results of research related to proven, effective, and evidence-based mathematics instruction. The Panel's assertions and recommendations, therefore, need to be grounded in the highest quality evidence available from scientific studies. The highest-quality evidence that is actually available on some key topics, however, may not be of sufficiently high quality to support confident conclusions. So that the Panel may be systematic in identifying the quality of evidence on which its assertions and recommendations are based, criteria such as the following will be applied during the preparation and review of the final report.

I. Background: Categories of Internal and External Validity

There are three broad categories into which one can categorize research and the corresponding claims based on that research. First, there is the highest-quality scientific evidence, based on such considerations as the quality of the design, the validity and reliability of measures, the size and diversity of subject samples, and similar considerations of *internal* (scientific rigor and soundness) and *external* validity (generalizability to different circumstances and students). Hypothesis testing, especially the active search for disconfirmation, is a hallmark of high-quality research (e.g., Lewin, 1951; Platt, 1964). Hence, the Panel's strongest confidence will be reserved for studies that test hypotheses, meet the highest methodological standards (internal validity), and have been replicated with diverse samples of students under conditions that warrant generalization (external validity).

In addition to reviewing the best scientific evidence, the Panel is also charged with considering promising or suggestive findings that should be the subject of future research. Promising or suggestive studies do not meet the highest standards of scientific evidence, but they represent sound, scientific research that needs to be further investigated or extended. For example, laboratory studies showing significant effects of "desirable difficulties" (i.e., difficulties produced by challenging to-be-learned material) or of repeated testing on long-term retention could be extended to actual classrooms or existing curricula (e.g., Bjork, 1994; Roediger & Karpicke, 2004; see Cook & Campbell, 1979). The final category corresponds to statements based on values, impressions, or weak evidence; these are essentially opinions as opposed to scientifically justified conclusions. Issues such as what constitutes algebra are matters of expert opinion rather than of scientific evidence.

II. Quantity, Quality, and Balance of Evidence

A. Strong Evidence

All of the applicable high-quality studies support a conclusion (statistically significant individual effects, significant positive mean effect size, or equivalent consistent positive findings), and they include at least three independent studies with different relevant samples and settings, *or* one large high-quality multi-site study. Any applicable studies of less than high quality show either a preponderance of evidence consistent with the high-quality studies (e.g., mean positive effect size) or such methodological weaknesses that they do not provide credible contrary evidence. Factors such as error variance and measurement sensitivity clearly influence the number of studies needed to support a conclusion (reflected in such statistics as p-rep, the probability of replicating an effect; Killeen, 2005); the number and balance of studies that are indicated above are, therefore, rules of thumb (e.g., see evidence standards applied by the What Works Clearinghouse at www.whatworks.ed.gov).

B. Moderately Strong Evidence

Criteria for moderately strong evidence are the same as that for strong evidence, but with one of the following exceptions: there are only one or two high-quality studies, the effects have not been independently replicated by different researchers, or they do not involve different samples (i.e., diversity of characteristics) and settings.

C. Suggestive Evidence

One of the following:

a) There are some high-quality studies that support the conclusion (statistically significant effects, significant mean effects) but others that do not (nonsignificant). Those that do not are null, not negative (nonsignificant effect or mean effects, but not a significant negative effect). Any applicable moderate quality studies show a comparable pattern or better.

b) There are no high-quality studies, but all the applicable moderate-quality studies support the conclusion (statistically significant individual effects, significant positive mean effect size, or equivalent consistent positive findings) and there are at least three such studies.

D. Inconsistent Evidence

The evaluation of mixed evidence depends crucially on the quality of the designs and methods of each study. The results of high-quality designs trump inconsistent or null results of low-quality designs. Mixed results of high and/or moderate quality studies that are not consistent enough to fall into any of the previously described categories, and cannot be adjudicated by methodological criteria, are inconclusive.

E. Weak Evidence

Evidence is considered weak when only low-quality studies are available.

III. Applying the Criteria

To apply such criteria, each study on which an assertion or recommendation is based must be characterized as "high quality," "moderate quality," or "low quality." The standards for those designations will necessarily differ for the different kinds of research that are applicable to different issues and inferences (Shavelson & Towne, 2002). The primary interest of the Panel is experimental and quasi-experimental research designed to investigate the effects of programs, practices, and approaches on students' mathematics learning and achievement. On some matters, however, the relevant studies are surveys (e.g., of students' mathematical knowledge). On yet other matters, by necessity, the relevant sources represent compilations of practice and informed opinion (e.g., the mathematical concepts essential to algebra). The methodological quality of individual studies will be categorized as part of the documentation for the database for the Panel's work, using such definitions as the following.

For studies of the effects of interventions:

High quality. Random assignment to conditions; low attrition; valid and reliable measures.

Moderate quality. Nonrandom assignment to conditions with matching, statistical controls, or a demonstration of baseline equivalence on important variables; low attrition or evidence that attrition effects are small; valid and reliable measures. Correlational modeling with instrumental variables and strong statistical controls. Random assignment studies with high attrition.

Low quality. Nonrandom assignment without matching or statistical controls. Pre-post studies. Correlational modeling without strong statistical controls. Quasi-experimental studies with high attrition.

For descriptive surveys of population characteristics:

High quality. Probability sampling of a defined population; low nonresponse rate or evidence that nonresponse is not biasing; large sample (achieved sample size gives adequate error of estimate for the study purposes); valid and reliable measures.

Moderate quality. Purposive sampling from a defined population; face valid for representativeness; low nonresponse rate; moderate to large sample size; valid and reliable measures. Probability sample with high nonresponse rate, but evidence that nonresponse is not biasing.

Low quality. Convenience sample; high nonresponse rate or evidence that it is biasing; small sample size; invalid or unreliable measures.

For studies of tests and assessments:

Psychometric standards such as measures of validity, reliability, and sensitivity will be used to evaluate tests and assessments (e.g., Anastasi, 1968; Cronbach & Meehl, 1955).

IV. Task Group Guidelines

To ensure identification of the best available evidence in the research literature, each task group has developed guidelines for the literature search that identify the relevant topics and the screening criteria to be used to select the studies the task group will consider for review. These criteria are designed to produce full or representative coverage of the highest quality and the most relevant studies in a relatively efficient manner.

A. Learning Processes Task Group

1. Topics and content

a) Research linking mathematical content and children's learning, and cognitive processes. Focus on children's solving or understanding of mathematics in specific content areas (see key words) with measures of children's learning, problem solving, or understanding that are more precisely defined than is typically found with achievement measures, e.g., trial-by-trial assessment of problem solving strategy.

2. Coverage

a) Emphasis on the literature found in a designated set of core journals supplemented with studies on specific topics of interest (e.g., whole number division) from other peer-reviewed journals.
b) Reviews of empirical research in books or annual reviews (e.g., *Annual Review of Psychology*, *Handbook of Child Psychology*).
c) Published in English, 1990 or after; supplemented with earlier, high-citation impact work, where available.

3. Study samples

a) Children 3 years of age to young adult.

4. Study methods

a) Randomized experiments.
b) Quasi-experiments with nonrandom assignment to conditions.
c) Correlational studies with a measure of math processes that is predicted by or predicts some other achievement outcome or process measures.

B. Conceptual Knowledge and Skills Task Group

1. Topics and content

a) Topics taught and assessed in mathematics, preschool to eighth grade and algebra, in the United States and internationally.
b) The relationship between math concepts and skills learned or taught at elementary and middle school levels, and later success in algebra (achievement).

2. Coverage

a) State and international curriculum frameworks for preschool to Grade 8 mathematics topics.
b) Course-Level Expectations in State-based Curriculum Frameworks for the algebra topics [synthesized by Institute for Defense Analyses/ Science and Technology Policy Institute (STPI) for 22 states].
c) Contents of algebra textbooks with particular attention to current and historic (1913) algebra topics (synthesized by STPI for 27 textbooks).
d) Pre-algebra (kindergarten through eighth grade) and algebra topics represented in the National Assessment of Educational Progress (NAEP), the Advanced Diploma Project (ADP), and the Singapore Curriculum.

3. Study samples

a) Students from elementary through high school grades.

4. Study methods

a) Descriptive (frequency) analysis from representative sets of materials nationally and internationally.
b) Criteria established by manuscript authors (e.g., Fordham report) for state mathematics frameworks.

C. Instructional Practices Task Group

1. Topics and content

a) Effects of instructional practice, teaching strategies, and instructional materials on mathematics achievement.

2. Coverage

a) Published in a peer-reviewed journal or government report.
b) Published in English, 1976 or after.

3. Study samples

a) Children, kindergarten through high school level.

4. Study methods

a) Randomized experiments or quasi-experiments with techniques to control for bias (matching, statistical control) or demonstration of initial equivalence on important pretest variables.
b) Attrition of less than 20% or evidence that the remaining sample is equivalent to the original sample on important variables.

D. Teachers Task Group

1. Topics and content

a) Relationship between teacher content knowledge and student achievement.
b) Programs of teacher education and professional development, and their effects on teacher knowledge, instructional practice, and student achievement.
c) Programs of mathematics specialist teachers at the elementary level, and effects on instruction and student achievement.
d) Programs to recruit and retain qualified teachers, and their effects on teacher quality.

2. Coverage

a) Published in a peer-reviewed journal or government report.
b) Books and book chapters.
c) Selected reports relevant to key topics.
d) Published in English.

3. Study samples

a) Teachers of preschool through high school students.

4. Study methods

a) Randomized experiments.
b) Quasi-experiments with techniques to control for bias (e.g., matching, statistical control) or demonstration of initial equivalence.
c) Correlational studies of natural variation with statistical controls.

V. Procedures

A. Screening Criteria for the Literature Search

As described in the previous section, each task group has developed specific criteria for identifying and screening the research literature pertinent to its task. Those criteria give priority to high-quality scientific research but also include weaker evidence where it may be promising or suggestive, and when limited high-quality research is available. As such, the search and screening criteria do not provide an assessment of methodological quality per se; they only describe the studies each task group wishes to consider in preparing its review.

B. Documenting the Quality of the Evidence Used in the Report

The individual research studies that are considered part of the relevant research base by each task group will be evaluated as presenting high-, moderate-, or low-quality scientific evidence using the standards appropriate to the nature of the research. For some task groups, this coding will be done by Abt Associates, Inc. as part of their documentation of the database of research studies on which the Panel's review is based. The body of research on which each significant claim, conclusion, and recommendation in the report is based will be characterized as strong, suggestive, or weak according to the quality, quantity, and generalizability of the collective evidence across studies. This information will guide the wording of the Final Report with regard to the confidence with which conclusions and recommendations are presented.

VI. Recommendations

The Panel's systematic reviews have yielded hundreds of studies on important topics, but only a small proportion of those studies have met methodological standards. Most studies have failed to meet standards of quality because they do not permit strong inferences about causation or causal mechanisms (Mosteller & Boruch, 2002; Platt, 1964). Many studies rely on self-report, introspection about what has been learned or about learning processes, and open-ended interviewing techniques, despite well-known limitations of such methods (e.g., Brainerd, 1973; Nisbett & Ross, 1980; Woodworth, 1948). Therefore, the Subcommittee on Standards of Evidence recommends that the rigor and amount of coursework in statistics and experimental design be increased in graduate training in education. Such knowledge is essential to produce and to evaluate scientific research in crucial areas of national need, including mathematics education.

BIBLIOGRAPHY

Anastasi, A. (1968). Psychological testing (3rd ed.). Collier-Macmillan Ltd, London.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp.185-205). Cambridge, MA: MIT Press.

Brainerd, C. J. (1973). Judgments and explanations as criteria for the presence of cognitive structures. *Psychological Bulletin, 79*, 172-179.

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

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345-353.

Lewin, K. (1951). Field theory in social science: Selected theoretical papers. New York: Harper & Row.

Mosteller, F., & Boruch, R. (Eds.). (2002). Evidence matters: Randomized trials in education research. Washington, DC: Brookings Institution Press.

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, NJ: Prentice-Hall.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347-353.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.

Shavelson, R. J., & Towne, L. (2002). Scientific research in education. Washington, DC: National Academy Press.

Woodworth, R. S. (1948). Contemporary schools of psychology (2nd ed.). New York: Ronald Press.