

# What Doesn't Work: The Challenge and Failure of the What Works Clearinghouse to Conduct Meaningful Reviews of Studies of Mathematics Curricula

by Alan H. Schoenfeld

An early version of this article, discussing curricular interventions in mathematics, was written for the What Works Clearinghouse (WWC). The Institute of Education Sciences (IES), which funds WWC, instructed WWC not to publish it. An expanded version, written at WWC's invitation for a special issue of an independent electronic journal and a book to be published by WWC, argued that methodological problems rendered some WWC mathematics reports potentially misleading and/or uninterpretable. IES instructed WWC staff not to publish their chapters—thus canceling the publication of the special issue and the book. Those actions, chronicled here, raise important issues concerning the role of federal agencies and their contracting organizations in suppressing scientific research that casts doubt on current or intended federal policy.

---

**T**he What Works Clearinghouse was established by the U.S. Department of Education's Institute of Education Sciences (IES) in an attempt to do for education what "evidence-based health care" has done for the health professions.<sup>1</sup> The home page of the Clearinghouse website (<http://www.whatworks.ed.gov/>) describes the effort as follows:

On an ongoing basis, the What Works Clearinghouse (WWC) collects, screens, and identifies studies of the effectiveness of educational interventions (programs, products, practices, and policies). We review the studies that have the strongest design, and report on the strengths and weaknesses of those studies against the WWC Evidence Standards so that you know what the best scientific evidence has to say.

WWC does not conduct research. It scans the extant literature in search of studies that meet its very stringent methodological criteria, examines those studies, and reports on their findings. Ultimately, WWC's goal is to conduct meta-analyses to determine the effects of educational and other interventions. Studies that qualify for vetting by WWC must be of one of the following three types:

- Randomized experiments
- Quasi-experiments that use equating procedures
- Studies that use the regression discontinuity design

WWC subjects studies of these types to careful technical evaluations along multiple dimensions. Ultimately, a study may be

given one of three ratings: "meets the standard," "meets the standard with reservations," or "does not meet the standard." (For example, a randomized controlled experiment might fail to meet the standard if there is too large a difference in attrition rates between the experimental and control groups.) Informally, meeting the WWC standard is portrayed as being statistically akin to receiving the *Good Housekeeping* seal of approval. Although WWC does not endorse curricula, the intention is to provide consumers with information about successful interventions. If a study that meets WWC standards shows that a curriculum works—that is, the study provides evidence that a curriculum produces better student performance than a control treatment—the presumption is that educational consumers should have confidence that the curriculum in question is effective.

I was asked to join WWC as "senior content advisor" for its studies of mathematics curricula—first for the review of studies of middle school mathematics curricula, and later for mathematics curriculum studies at all levels. I had some misgivings, because the WWC agenda is very narrow; I believe that many factors other than those on the WWC agenda should be taken into account when examining curricular effectiveness (see National Research Council [NRC], 2005, for an extensive discussion of this issue). However, I also believe that properly conceived and executed comparative studies can contribute to our understanding in important ways. Some of the issues involved are subtle, and it is easy to go astray. Thus, when offered the opportunity to pitch in, I agreed to do so. I felt a moral and intellectual responsibility to help make sure that things would be done well.

One of my first tasks as senior content advisor was to help put together the protocol to be followed by WWC staff in evaluating middle school mathematics curricular interventions. That document was chock full of technical detail, with more than 50 pages of information regarding criteria for the WWC literature search; a list of databases and journals to be searched; a list of keywords for electronic searches; procedures for collecting "fugitive" or "grey" literature; a list of organizations and publishers to contact about sources of curricular interventions; methods for classifying studies; methods for coding studies; criteria for determining the independence of multiple outcome measures; statistical procedures and conventions for conducting meta-analyses; rules for computing effect sizes; sensitivity analysis; power analysis; and more. Although I vetted much of the technical detail, my central contribution was an essay on "background and context." The purpose of the section containing my essay was to explain why it is necessary to conduct curricular evaluations in the first place and what some

of the major considerations in doing so should be. The core ideas from that essay, which relate to curriculum assessment and epistemological issues, are presented here, in the following section.

### Issues in the Evaluation of Mathematics Curricula

Mathematics has been taught and assessed in the United States for more than a century. A variety of measures, including the National Assessment of Educational Progress (NAEP), standardized tests such as the ACT and SAT, and more recent measures such as the New Standards and Balanced Assessment examinations, have been employed to assess student competencies in mathematics. It is reasonable to assume, then, that myriad data would be readily at hand to address the question on which my work at WWC was focused: *How effective are various curricula in helping students to learn mathematics?*

The situation is not that simple. This article, which includes a very brief description of the curriculum assessment context in the United States, explains why. (For more extended discussions of the evolution of American mathematics curricula over the 20th century, and the forces behind recent curricular changes, see Schoenfeld, 2001, 2002; see also Begle, 1970; Henry, 1951; *Journal for Research in Mathematics Education*, 1994; Kilpatrick, 1992; National Council of Teachers of Mathematics [NCTM], 1989, 2000; Whipple, 1930.)

In what follows, I address some of the complexities that are involved in the ostensibly simple task of mathematics curriculum evaluations. In broad-brush terms, I provide some historical background, a summary of current controversies, and a discussion of measures and data that can be used to work through those controversies.

#### *Historical Background*

What is the historical context for the assessment of mathematics curricula? Are there differences of opinion regarding “what counts” in mathematics instruction?

American mathematics curricula have periodically become the focus of intense national attention. Typically, this has happened in times of war or other perceived national crises, such as the launch of the Soviet satellite Sputnik during the Cold War. In the late 1970s and 1980s, the crisis that stimulated a reexamination of American mathematics and science education was economic. *A Nation at Risk* (National Commission on Excellence in Education, 1983) summarized the situation as follows:

If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war. We have, in effect, been committing an act of unthinking, unilateral educational disarmament. (p. 1)

Alongside the economic crisis was evidence of American students' poor mathematical performance on the Second International Mathematics Study (McKnight, Crosswhite, Dossey, Kifer, Swafford, Travers, & Cooney, 1987; McKnight, Travers, & Dossey, 1985). Thus the stage was set for curricular change.

The idea of addressing a mathematical crisis by producing new curricula is not new. In the 1960s, for example, a range of curricula in mathematics and the sciences was created in response to the Sputnik launch. These included curricula known collectively as

the “new math.” The new mathematics curricula broke with a traditional emphasis on skills and focused on aspects of conceptual understanding, in a way that was unfamiliar to many parents and teachers. They were controversial and ultimately sparked a strong counter-movement, called “back to basics.” Curricula focusing on skills dominated much of the 1970s. Thus curricular history provides prior examples of the tension between “teaching for skills” and “teaching for understanding” that is hotly debated today. There is a difference, however. The “reform” curricular materials developed in the 1990s had a significantly different grounding in research than the curricula that preceded them. The cognitive revolution (Gardner, 1985; see also De Corte, Greer, & Verschaffel, 1996; Lester, 1994; Schoenfeld, 1985, 1992) had fundamentally changed the research community's understanding of mathematical thinking and learning.

As detailed below, research in the 1970s and 1980s resulted in a reformulation of the very notion of competency in mathematics and other subject areas. In brief, the shift was from an emphasis on what students know to an emphasis on the knowledge and processes that enable students to use their mathematical knowledge effectively—colloquially, to an emphasis on what students “know and can do.” Partly as a result of this research, goals for mathematics instruction expanded. Research on mathematical thinking and problem solving, along with an expanded set of social goals, constituted the intellectual underpinnings of an extraordinarily influential document entitled *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989). The *Standards*, as they are known, called for a broad set of goals for mathematics instruction (including, for example, mathematical literacy as a component of literate citizenship and of preparation for the workplace, as well as preparation for more advanced mathematics). In line with contemporary research, the *Standards* emphasized mathematical processes such as problem solving and reasoning, making connections, and communicating with mathematics.

The *Standards* were not prescriptive—they described desired outcomes but left open the specific curricular ways of achieving those outcomes. There was room for wide interpretation regarding how to achieve the goals set forth in them. Over the years following the publication of the *Standards* in 1989, a number of very different “reform” or “standards-based” curricula were developed, many with major funding from the National Science Foundation. Most differed in significant ways from traditional curricula. Some focused heavily on applications; some made extensive use of group work; some used “manipulatives” and other “hands-on” materials. Collectively, they sparked significant controversy, known as the “math wars.” (These are described later; see also Schoenfeld, 2004.)

Here is the current situation with regard to our understanding of the effectiveness of mathematics curricula. The Third International Mathematics and Science Study, known as TIMSS, was conducted in six grades in more than 40 countries in 1994–1995. The TIMSS comparative data (see, e.g., Beaton, Mullis, Martin, Kelly, & Smith, 1996), indicated that the mathematical performance of American students was far lower than desirable. The TIMSS curricular reports suggested that at least part of the problem resided in American curricula, which were seen as more skills oriented, more repetitive, and less conceptually deep than those of nations that scored better on TIMSS (Schmidt, McKnight, & Raizen, 1997). A new generation of “standards-based” curricula now exists. These curricula,

in line with contemporary research on mathematical thinking and problem solving, tend to have much broader goals than did the prior, “traditional” curricula; they also focus less on skills and symbol manipulation (see, e.g., Nathan, Long, & Alibali, 2002). They are controversial, having both advocates and detractors.

The fact is that these “reform” curricula have just recently entered the mainstream. Generally speaking, polished versions of the new curricula became available in the middle and late 1990s. Relatively small numbers of students have worked their way through a full reform curriculum. In consequence, there are scant data regarding the effectiveness of these curricula—either on their own merits or in comparison with traditional curricula. On the positive side, the evaluations of reform curricula do tend to take into account the broad set of mathematical performance goals, including problem solving, that are deemed central by current research. Thus assessments of these curricula do tend to capture at least three relevant aspects of student performance: the knowledge base, conceptual understanding, and problem solving.<sup>2</sup> Ironically, comparable data are extremely rare for the traditional curriculum, despite its near-universality for many years. Traditional measures such as NAEP have not been used to assess traditional curricula and are not appropriate to that purpose. The dimensions of mathematical performance deemed central by contemporary research were not explicitly highlighted or measured as the traditional curriculum evolved and stabilized. Hence, generally speaking, exemplars of the traditional curriculum have been examined along those dimensions only when being compared with “reform” curricula.

In sum, there are no definitive findings regarding the effectiveness of either traditional or reform curricula that take into account the spectrum of mathematical competencies that are now understood to be central to the effective understanding and use of mathematics.

### *Current Controversy*

Do debates exist about the utility of various interventions? Are there different theoretical or empirical stances regarding relevant approaches and their impact?

The answer to these questions can be summed up in two words: “math wars.” In recent years mathematics curricula have been a topic of significant controversy. What is typically referred to as the traditional mathematics curriculum has been in place, to varying degrees, since the 1950s.<sup>3</sup> By and large, the focus of the traditional curriculum was on mathematical knowledge—the mastery of procedures and the ability to demonstrate understanding by solving problems. Thus, for example, when a minor curricular reform in the early 1980s emphasized “problem solving,” it meant that some of the emphasis in solving arithmetic problems at the elementary grades changed from computations such as

$$7 - 4 = \underline{\quad}$$

to “word problems” such as

*John has 7 apples. He gives 4 apples to Jane.  
How many apples does John have left?*

However, even “two-step” problems at later grades such as

*Mary bought a box of tissues for \$1.49 and a pen for \$0.79.  
She paid with a \$5.00 bill. What change did she receive?*

were considered to be difficult and nontraditional. NAEP results, as well as numerous research studies, showed that students had difficulties with such problems. Generally speaking, curricula and assessments focused on the mastery of core concepts and procedures. For a topic area such as fractions, for example, in typical assessment tasks students would be asked to (a) add, subtract, multiply, and divide two fractions; (b) find a third fraction whose value lay between two given fractions; (c) estimate the result of some operation with fractions; or (d) convert fractions to decimals, or vice-versa.

The prevalence of the traditional curriculum waxed and waned periodically over the latter half of the 20th century, often in reaction to perceived crises in mathematics education. As noted above, the “new math” arose in the wake of the Soviet Union’s launch of the first earth-orbiting satellite, Sputnik; it was swept aside in the “back to basics” movement of the 1970s. The “problem solving” movement of the 1980s arose partly in response to the realization that student mastery of the basics had not significantly improved after a decade of emphasis on core skills. Thus there have been changes over time—but core aspects of the traditional curriculum remain in place. Variants of the traditional curriculum (represented by the textbook series of the major publishers) remain dominant to this day.

The strongest challenge to the domination of the traditional curriculum has come, in the past decade, from what are generally called reform or standards-based mathematics curricula. To understand what they represent, and the assessment challenges they pose (both on their own terms and in contrast to the traditional curriculum), requires an understanding of significant changes in researchers’ understanding of the nature of mathematical thinking since the 1970s.

For most of the 20th century, the dominant perspective on learning in most fields, and specifically in mathematics, was that learning is the accumulation of knowledge; that practice solidifies mastery; and that knowledge is demonstrated by the ability to solve particular (well-studied) classes of problems. Over the 1970s and 1980s, evidence mounted that a wide range of other skills and understandings were central to effective mathematical performance. Those included having a solid knowledge base, much as would be expected in the traditional curriculum; being able to employ a range of problem-solving strategies; being able to reason effectively using mathematical ideas and to communicate one’s reasoning effectively, orally and in writing; being able to make effective use of various resources, including the knowledge and time at one’s disposal; and having a productive set of beliefs and dispositions about the nature of the mathematical enterprise (see, e.g., De Corte, Greer, & Verschaffel, 1996; Schoenfeld, 1985, 1992).

More recent volumes span the same space. For example, the NRC’s volume *Adding It Up* (2001, p. 5) describes five interwoven strands of mathematical proficiency:

- *Conceptual understanding*—comprehension of mathematical concepts, operations, and relations
- *Procedural fluency*—skill in carrying out procedures flexibly, accurately, efficiently, and appropriately
- *Strategic competence*—ability to formulate, represent, and solve mathematical problems
- *Adaptive reasoning*—capacity for logical thought, reflection, explanation, and justification

- *Productive disposition*—habitual inclination to see mathematics as sensible, useful and worthwhile, coupled with a belief in diligence and one’s own efficacy


More fine-grained analyses of proficiency tend to be aligned with the content and process delineations found in NCTM’s (2000) *Principles and Standards for School Mathematics*:


*Content:* Number and Operations, Algebra, Geometry, Measurement, Data Analysis and Probability

*Process:* Problem Solving, Reasoning and Proof, Making Connections, Oral and Written Communication, Uses of Mathematical Representation

As noted earlier, traditional assessments look for students’ ability to compute with fractions (e.g., “Find  $\frac{2}{5} + \frac{3}{4}$ ”). More recent assessments, aligned with the broad characterizations of mathematical proficiency just listed, ask for more. See Figure 1 for examples of assessment tasks that have been used to judge various aspects of students’ knowledge of fractions.

Tasks 1 and 2 check for students’ conceptual understanding. They ask students to work with different ways of representing fractions, and they check to see whether students understand that a given fraction of the whole must be the same size in every occurrence. (Many students will answer “ $\frac{1}{4}$ ” for Task 1 and “ $\frac{2}{6}$ ” for Task 2.) Task 3 probes students’ understandings of proportional and inverse relationships (when the denominator increases, the value of the fraction decreases) as well as their ability to explain what they understand.

1.  Write a fraction for the shaded part of the region. \_\_\_\_\_  
Now write an equivalent fraction. \_\_\_\_\_

2.  Write a fraction for point A. \_\_\_\_\_  
Now write an equivalent fraction. \_\_\_\_\_

3. The figure below can stand for any fraction:

$$\frac{\text{numerator}}{\text{denominator}}$$

- If the numerator of the fraction is multiplied by 2 and the denominator stays the same, how does the value of the fraction change? Explain.
- If the denominator of the fraction is multiplied by 2 and the numerator stays the same, how does the value of the fraction change? Explain.

FIGURE 1. Assessment items that focus on a conceptual understanding of fractions.

Another recent assessment task gives students the ratio in which yellow and red paint are mixed to make orange paint and the ratio in which orange and blue paint are mixed to make brown paint. The students are then asked to determine the proportion of yellow paint in the brown paint, how many gallons of red paint will be needed to make so many gallons of brown paint, and so forth. Thus students can demonstrate the ability to use their knowledge of fractions in contexts other than formal mathematics. This problem and those presented in Figure 1 all capture aspects of “understanding fractions.”

The assumptions underlying the traditional curriculum have, of course, included the idea that the conceptual underpinnings of arithmetic procedures should be explained. To generalize somewhat, a core assumption underlying traditional curricula is that students should practice mathematical procedures until they have mastered them, and that deeper conceptual understanding and “transfer” to new situations will come with increasing fluency (see, e.g., Henry, 1951; Whipple, 1930).

The assumptions and practices supporting the reform curricula are more varied but show some consistencies as well. Many such curricula engage students in complex problems—sometimes before the “basics” have been mastered, sometimes as a way of providing contexts within which the basics can be learned.

In various ways, then, some of the reform curricula are designed to support students’ abilities to engage with tasks such as the fractions tasks discussed here. As such, they try to capture some aspects of mathematical understanding revealed by contemporary research. However, focusing on issues of mathematical representation and problem solving takes time. The cost of that focus is that the newer curricula allow less time for mastery of basic skills. The argument made by advocates of the traditional curriculum is that basic knowledge is a prerequisite for applications, and that students will be seriously hampered by their lack of foundational skills. The argument made by advocates of reform is that skills will develop in more robust fashion if they are developed in meaningful problem solving contexts.

This is somewhat contested territory. Generally speaking, advocates of reform decry the well-documented consequences of the traditional curricula through the 1990s: poor U.S. performance on international comparison studies, significant attrition rates in mathematics, which are attributed to the ostensibly unappealing nature of the curriculum; and racial performance gaps, with data showing that Latinos, African Americans, and Native Americans do much more poorly than Whites and some Asians. They argue that the new approaches are more enfranchising, more mathematically meaningful, and less likely to cause the documented problems of the traditional curriculum. Advocates of the traditional curriculum fear that the reform curricula have thrown out the mathematical baby with the bathwater—that the curricula rest on a series of unfounded assumptions about “transfer,” that they depend too much on discovery and faddish classroom practices such as group work, and that they do not establish the firm mathematical foundations required for successful applications of the mathematics. In my opinion, no truly robust comparisons of the impact of the two kinds of curricula have been done. Indeed, the very existence of these controversies was one of the motivations for the creation of WWC.

### *Working Through the Controversy*

What measures are appropriate, and what kinds of data are available, to resolve these issues?

Appropriate measures of mathematical performance are those that capture the dimensions of mathematical proficiency discussed in the previous section. Recall the framework from *Adding It Up* (NRC, 2001): conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition. In standard testing terms, the first three of these are typically referred to as concepts, skills, and problem solving. There are perhaps two widely accessible assessments that explicitly cover those aspects of mathematical proficiency: the Balanced Assessment examinations produced by CTB-McGraw Hill and the New Standards examinations produced by Harcourt Brace Educational Measurement. The fourth strand, “reasoning,” is often embedded in the performance items on these tests; “disposition” is typically assessed differently, either by questionnaire or observation. There do exist, then, neutral and independent tools that can be used to measure desired outcomes. However, these tools have not had generally wide use—especially for comparative purposes. Typically, studies of individual curricula have used standardized tests to measure procedural fluency, and one or more locally developed measures to capture aspects of conceptual understanding such as problem solving, reasoning, and disposition.

Although the traditional curriculum has existed for many years in various incarnations—as noted, mainstream textbook series were relatively close to each other in content—there exists very little by way of quantitative evaluation of individual curricula. Until the passage of the No Child Left Behind Act of 2001, publishers had little or no incentive to gather data regarding student performance, because the marketplace did not demand it (Burkhardt & Schoenfeld, 2003). Moreover, the existing quantitative studies are generally limited in terms of content, because of test design. Most standardized tests were designed under the standard psychometric assumptions of trait and/or behaviorist psychology (Glaser & Linn, 1997; Greeno, Pearson, & Schoenfeld, 1997), and they do not measure the varied aspects of mathematical proficiency described above.

This is not a trivial matter. Tests that focus on only a subset of the desired range of performance can give misleading results. Consider, for example, a study by Ridgway, Crust, Burkhardt, Wilcox, Fisher, and Foster (2000). The study compared students’ performance at Grades 3, 5, and 7 on a standardized high-stakes, skills-oriented test (the California STAR test) with their performance on a much broader standards-based test (the Balanced Assessment test). Scores on each test were divided into two simple categories: “proficient” or “not proficient.” The data indicated that between 70% and 75% of the students at each grade level scored equivalently (either proficient or not proficient) on both tests. However, fewer than 5% of the students scored proficient on standards-based test and not proficient on the skills-oriented test, while about 22% of the students were deemed proficient on the skills-oriented test but not proficient of the standards-based test. The latter group of students, nearly a quarter of the student population, was deemed “proficient” by the state of California on the basis of the STAR test, but that ostensible proficiency may well have been an artifact of the narrowness of the test. Those students’

low scores on the Balanced Assessment tests suggest that the “proficient” ratings on the STAR tests may be “false positives.” That is, the students’ proficiency is called into question when measures reflecting contemporary research are employed.

Similarly, there can be significant curricular “false negatives” when curricula are compared. Suppose that students learn much more from Curriculum A than from Curriculum B, but what they have learned over and above Curriculum B goes untested by a skills-oriented outcome measure used for comparative purposes. A randomized controlled trial using the skills-oriented measure to compare Curriculum A with Curriculum B would report “no significant differences” because it would be insensitive to significant differences in conceptual understanding and problem solving between the two treatments. In sum, curricula must be assessed according to all of the relevant criteria.

Given this situation, it is not at all clear how much of the extant literature can provide the information necessary to make meaningful comparisons of mathematics curricula; nor is it clear what kinds of information can result from syntheses of those studies. For the most part, traditional curricula, if they have been evaluated at all, have not been evaluated along the relevant dimensions of proficiency discussed earlier. The picture is not much prettier when one looks at the reform side of the curriculum ledger. True, the developers of newer curricula have felt the pressure to document the impact of their work. However, their products are “just off the shelf” in terms of usage; given their recent introduction, few cohorts of students have studied any particular reform curriculum from beginning to end. Moreover, many of the assessments used to measure the effectiveness of the new curricula were locally constructed. Only in the past few years have compendia of assessment results for these curricula (see, e.g., Senk & Thompson, 2003) or large-scale comparative studies (see, e.g., the ARC [Alternatives for Rebuilding Curricula] Center Tri-State Student Achievement Study, 2003) begun to provide quantitative data regarding the comparative performance of traditional and standards-based curricula.

There is a further complication: It is far from clear what it means to “implement” a curriculum. Some will argue that there is “no such thing” as a curriculum, per se: What matters is the character of the implementation in context. Indeed, one can imagine curricular materials that, when used in the way intended by the designers, result in significant increases in student performance, but, when used by teachers not invested or trained in the curriculum, result in significant decreases in student performance. Hence, data gathering, coding, and analysis must try to indicate the character of the implementation and its fidelity to intended practice. Ideally, results disaggregated by the conditions of implementation would provide evidence of what the curriculum’s effects might be in varied contexts—and, more important, evidence of what kinds of support structures are helpful in various implementation contexts (i.e., urban, rural, or suburban schools; schools with high proportions of second language learners, etc.). Undifferentiated results (or those in which the conditions of implementation are not known), over a large population, might be seen as approximating (the range of) “general” implementation of the curriculum. In the event that information on the quality of implementation is not known, the default assumption will be that the implementation

is “typical” and that differences in implementation will wash out statistically.

### The Fate of the Essay, Part I

The draft protocol containing my essay on background and context (essentially the preceding section) was submitted to IES for review in October 2003. It was returned by IES with instructions to remove the essay. I expressed my concern about this more than once to the WWC leadership and received the following e-mails by way of reassurance:

I just wanted to let you know that we actually loved the math background section, and want to make it the background of the bigger topic report. We’ve been having a lot of talks with ED [the U.S. Department of Education] lately about the target audience for each of our documents, to help us figure out what level to write at. For the proposal, we all agreed that this was more of a technical document for the review teams though it will eventually become a technical appendix to the topic report. (Stephanie Cronen, WWC, December 2, 2003)

We have every intention of presenting the full background in the topic report for math. ED didn’t seem to have any comments on the content, just told us to move the longer backgrounds to the topic reports. We’ve talked a lot about censoring within the project, however, and plan to proceed as if it will not occur. As long as everything we present is fair and balanced, there’s really nothing they can say to convince us otherwise. They know we’re sensitive to this already. (Stephanie Cronen, December 7, 2003)

### Issues in the Evaluation of WWC Study Reports

Over the course of the following year, WWC conducted its analyses of middle school curriculum evaluations. In late 2004, WWC issued its middle school mathematics topic report. My introductory essay was nowhere to be seen. However, its predictions regarding the paucity of extant studies in the literature meeting WWC’s methodological criteria were substantiated:

From a systematic search of published and unpublished research, the What Works Clearinghouse (WWC) identified 10 studies of 5 curriculum-based interventions for improving mathematics achievement for middle school students. These include all studies conducted in the past 20 years that met WWC standards for evidence. (What Works Clearinghouse, 2004a, p. 2)

Unfortunately, the WWC study reports raise some very serious issues. A recent *Education Week* article covered one of those reports (What Works Clearinghouse, 2004b) as follows:

James J. Baker, the developer of a middle school mathematics program known as Expert Mathematician, is also dismayed at the way his research on the program is reported. His study—the only one that fully met the criteria for this topic—used a random assignment strategy to test whether students could learn as much with his student-driven, computer-based program as they could from a traditional teacher-directed curriculum known as Transition Mathematics. The problem, he argues, is that the [WWC] web site said his program had no effect without explaining that students made learning gains in both groups. (Viadero, 2004, p. 32)

This kind of issue can be resolved by expanding the information provided in WWC reports. After all, it is a simple matter to report

gains for the curricula being compared in comparative studies. Another study report, however, can only be described as fundamentally flawed—precisely because it ignored the concern (discussed in the first part of this article) about the character of the assessments used to compare curricula. In that report (What Works Clearinghouse, 2004c), WWC determined that a study by C. Kerstyn (2001), a quasi-experimental design with matching, met the WWC standards “with reservations.” (The reasons for the reservations were concerns with regard to implementation fidelity, sampling characteristics, and which subgroups were tested.) However, the statistical analyses in the study got full marks. The report says:

The fifth outcome is the Florida Comprehensive Assessment Test (FCAT), which was administered in February 2001. The author does not present the reliability information for this test; however, this information is available in a technical report written by the Florida Department of Education (2002). This WWC Study Report focuses only on the FCAT measures, because this assessment was taken by all students and is the only assessment with independently documented reliability and validity information. (p. 4)

Note that the report includes no independent examination of the content of the Florida Comprehensive Assessment Test (FCAT). Statistical tests of reliability and validity merely assess the psychometric properties of an examination; they do not provide information on whether, for example, the test covers the full spectrum of mathematical content that contemporary research declares to be essential. Until such an independent analysis is conducted—perhaps by experimental comparison with broad spectrum tests such as the Balanced Assessment or New Standards tests, perhaps by having an independent group examine the FCAT against a set of standards analogous to those in the NRC’s *Adding It Up* or in the NCTM’s *Principles and Standards for School Mathematics*—it is entirely possible that the FCAT, like the STAR test, is not an appropriate measure of mathematical performance.

Whether or not this is the case for the FCAT, the possibility represents a very serious problem in general. Suppose that, like the STAR test, the examination used as the outcome measure in an experimental study covers only some of the dimensions of the mathematics content and processes that contemporary research says should be taught and tested. In that case, as noted above, the outcome measure would be unable to distinguish between students who had demonstrated proficiency on a limited subset of mathematical understandings and students who could demonstrate proficiency on the broad spectrum of mathematical competencies deemed important by contemporary research. Such a measure might result in “false positives” at the individual student level, in the way that the STAR test did. Equally important, it might result in “false negatives” in the evaluation of curricula.

The danger of reporting false negatives is not merely hypothetical. One of the general arguments in favor of reform curricula (supported by all of the studies reported in Senk & Thompson, 2003) is that reform curricula do “about the same” as traditional curricula on measures of skills but much better on measures of conceptual understanding and problem solving. Suppose this is the case. If the outcome measure that is used in a comparison study focuses only on skills, it will report “no differences.” If the outcome measure examines skills, concepts, and problem solving,

then it will report differences favoring the experimental treatment. Different measures will produce different results. Thus the fact that an outcome measure has been shown to be statistically reliable and valid is not enough—one must know precisely what the test covers. Without that information, it is impossible to interpret the findings of the study. Equally important in terms of WWC’s mission, there will also be fundamental problems with the aggregation of data over different studies. To conduct a meaningful meta-analysis of studies that examine a particular curriculum, one must know the content that was being assessed in each study. A failure to conduct content analyses of the outcome measures used in comparative studies undermines the very purpose for which WWC was created.<sup>4</sup>

Although the focus of this essay has been mathematics, it should be noted that the issues discussed here apply to curriculum assessments in *all* content areas. As was noted by one of the reviewers of a draft of this article, many of the researchers who have objected to the summary report of the National Reading Panel (National Institute of Child Health and Human Development, 2000) could make similar arguments. Some measures of “reading comprehension” assess the ability to make sense of single sentences in isolation. Others assess the ability to read a paragraph or two and draw inferences from what one has read. The differences are consequential. Similarly, the need to focus both on broad mathematical/scientific content and on an understanding of cognitive processes as part of curriculum assessment lies at the core of work of the American Association for the Advancement of Science (AAAS; see, e.g., its Project 2061 publication *Benchmarks for Science Literacy*, 1993) and the series of text evaluations conducted by AAAS that can be accessed at <http://www.project2061.org/publications/textbook/default.htm>.

## The Fate of the Essay, Part II

In mid-2004 I was invited by the WWC staff member with whom I had worked on the WWC middle school mathematics studies to retool my introductory essay for inclusion in a special issue of the electronic journal *Research in Middle Level Education (RMLE) Online*. The special issue was to be sponsored by WWC and devoted to its work. The issue would consist largely of pieces written by WWC personnel, supplemented by my essay and an article by Jere Confrey, who had chaired the 2005 NRC panel on the evaluation of mathematics curriculum assessments. WWC was also considering publishing a hardcopy version of the special issue. I agreed, saying that I needed to re-do my paper because some time had passed and WWC had published some mathematics study reports in the interim. To the content of my first essay I added the core content of the preceding section of this paper.

The papers intended for the special issue made for an interesting and lively collection. Confrey’s paper was critical of the WWC enterprise, challenging the philosophical basis of the work and some of the technical decisions made by WWC. Mine took the WWC paradigm as a given, and asked what needed to be done for the analyses be useful and informative. This was, I thought, a paradigmatic example of open scholarly exchange.

When I submitted the final version of my contribution, I e-mailed the editor of the special issue, Stephane Baldi of WWC, asking when the final version of the manuscript would appear. He e-mailed me the following on February 18, 2005:

Right now it’s slotted for sometimes in April, but I will confirm once I have a definitive date. Also, the Dept. needs to “sign off” on the issue, so hopefully that won’t delay anything.

I think we ended up with a series of very interesting paper and we’re looking into possibly releasing it as an edited volume as well under AIR [American Institutes for Research]. Will let you know if this happens.

The same day, I e-mailed the following in reply:

For the record, I don’t see why the Dept. has to “sign off” on this or any academic product. NSF [the National Science Foundation] never does; we just put the official NSF disclaimer on articles. It’s something like this: “The author thanks NSF for its support under grant XXX. The foundation is not responsible for the opinions expressed herein.” (or words to that effect, I didn’t look up the official disclaimer.) That’s standard policy. In fact, the notion that NSF or any other federal agency might exert some form of censorship raises pretty serious issues of academic freedom.

His response:

I totally agree with you in terms of the “sign off” issue. However, Becki [WWC Project Director Rebecca Herman] indicated that since the project was brought together on their dime, they have the rights to review it before it goes to press. Hopefully, they’ll just do a passive review/sign off. (February 22, 2005, e-mail)

And my return e-mail the same day:

Let’s hope you’re right, and that there will be a passive sign-off. As a matter of courtesy, the Dept had a right to look at the papers—though NSF has never asked for that right *before* publication, we include copies of papers with annual reports. But if Department “review” means trying to compel any changes, that’s another matter altogether. It would be a clear case of federal censorship.

Then, on March 9, I received the following e-mail from Stephane Baldi:

I’m terribly sorry to have to convey this news to you but the U.S. Department of Education, our client on WWC, has instructed AIR that it must withdraw from the RMLE special issue. After having reviewed all the papers, their argument is that the papers do not reflect the current direction of the What Works Clearinghouse because they are based on a model that is out of date (i.e., the old model of study reports and coding guides based on the DIAD discussed in the papers). As contractors to the Department, we have no choice but to pull out. However, because you are not an AIR employee, you are still free to go ahead and publish your paper in whatever outlet you choose (including RMLE).

The IES rationale for forcing WWC to withdraw the papers does not stand up to examination. The reports that are currently posted on the WWC website were produced according to the procedures described in those papers, so the issues discussed *are* current. The only plausible interpretation for the withdrawal of the volume is that it provided a context for strong intellectual critiques by Confrey and me, and that IES did not want the special issue to serve as a mechanism for making our critical comments public. In my case, it was the third time that WWC/IES broke a commitment to make the ideas public. This cannot be viewed as anything other than an attempt to suppress the expression of scholarly work that calls into question some aspects of the WWC enterprise.

After giving the issue extended thought, I wrote the following e-mail to Stephane Baldi and WWC Project Director Rebecca Herman:

My personal opinion is that you should appeal this decision, to the highest levels, at IES. The decision to kill the volume will be seen as a clear act of censorship by IES and cowardice by WWC, especially since Jere's article and mine contained some strong criticisms of the current state of the enterprise. (March 19, 2005)

A week later, I had heard nothing in response. I submitted my resignation, with the following comments:

At this point I am forced to conclude that WWC is at minimum complicit in an act of censorship, having acceded from the beginning to IES requests to remove important material from its reports; that WWC's promises to make the work visible while acting to remove it were simply a form of stonewalling; and that WWC's refusal to deal directly with the fundamental issue of content analyses in its published reports renders its findings of little or no value and subject to serious misinterpretation. Much to my dismay—I had hopes for this enterprise and have put substantial energy into it—I am forced to conclude that the enterprise as currently constituted has neither intellectual nor moral integrity. (March 27, 2005, e-mail and attachment on letterhead)

Two days later I received the following message:

I'm sorry that you are resigning from the WWC, and also sorry for the frustration that contributed to this decision. This has been (and continues to be) a complex and challenging project, and I'm sorry that things haven't worked out the way you would have liked. I wish it had worked out differently.

While I don't think it would be productive to revisit every issue in your letter, I would like to point out that you are free to publish the article you wrote for RMLE. Below is the email we sent to RMLE making that point, which I believe we also conveyed to you. [Omitted.] So I don't think the decision on RMLE constitutes censorship of your article. (Rebecca Herman, March 29, 2005, e-mail)

I will leave it to readers to interpret that statement in light of the history given here. I will also leave it to readers to draw parallels between this and other attempts by federal agencies, past and present, to make it difficult for information that is contrary to current ideology to see the light of day. I will note that I find the entire story rather sad. I believed when I signed up for my stint at WWC, and I still do, that properly conducted quantitative research has an important contribution to make, as one of many ways to explore the impact of educational interventions (see, e.g., Schoenfeld, in press). But unless and until WWC changes its ways, and IES supports and encourages the open exchange of ideas, it will not make that contribution. We are poorer for that loss.

## NOTES

I am grateful to the reviewers and editors for their comments on a draft of this article. The revised version is stronger as a result.

<sup>1</sup>Evidence-based medicine was pioneered by the Cochrane Collaboration. A description can be found on the Cochrane Collaboration website, <http://www.cochrane.org/index0.htm>.

<sup>2</sup>As discussed in the next section, this is a minimal set of dimensions for assessment.

<sup>3</sup>"To varying degrees" is an important caveat. The United States has no national curriculum; states maintain the right to establish their own curricular goals and assessments. Until the past decade, the approximately 15,000 school districts in the United States had significant autonomy in choosing how to meet those goals (e.g., adjacent districts could use different curricula and even have different goals), and there were few statewide assessments. However, a substantial amount of homogeneity was established by the relative uniformity of national textbook series. Textbook series were produced in nationwide editions. To avoid losing a large potential market share, textbook producers were careful to meet the desiderata of three large states—California, New York, and Texas—each of which had established statewide curricular goals and textbook adoption criteria. Research shows that teachers tended to stay fairly close to the texts in coverage. Thus there was a de facto national curriculum of sorts, despite the ostensible latitude possessed by school districts in text selection and instruction.

<sup>4</sup>It is important to note that I have raised this issue in various ways with WWC staff. Despite my urging, and despite numerous promises that the issue would be given serious consideration, it was not brought by the staff to the WWC's Technical Advisory Group or taken up in any serious way.

## REFERENCES

- Alternatives for Rebuilding Curricula Center. (2003). *Tri-State Student Achievement Study*. Lexington, MA: Author. Available at <http://www.comap.com/elementary/projects/arcl/index.htm>
- American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. Washington, DC: Author.
- Beaton, A., Mullis, I., Martin, M., Kelly, D., & Smith, T. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Boston: International Association for the Evaluation of Educational Achievement.
- Begle, E. G. (Ed.). (1970). *Mathematics education* (69th yearbook of the National Society for the Study of Education). Chicago: National Society for the Study of Education.
- 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.
- De Corte, E., Greer, B., & Verschaffel, L. (1996). Mathematics teaching and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 491–549). New York: Macmillan.
- Gardner, H. E. (1985). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.
- Glaser, R., & Linn, R. (1997). *Assessment in transition: Monitoring the nation's educational progress: Background studies*. Stanford, CA: National Academy of Education.
- Greeno, J. G., Pearson, P. D., & Schoenfeld, A. H. (1997). Implications for the National Assessment of Educational Progress of research on learning and cognition. In *Assessment in transition: Monitoring the nation's educational progress: Background studies* (pp. 152–215). Stanford, CA: National Academy of Education.
- Henry, N. (Ed.). (1951). *The teaching of arithmetic* (50th yearbook of the National Society for the Study of Education, Part II). Chicago: University of Chicago Press.
- Journal for Research in Mathematics Education*. (1994). 25th anniversary special issue. Volume 25, No. 6.
- Kilpatrick, J. (1992). A history of research in mathematics education. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 3–38). New York: Macmillan.
- Lester, F. (1994). Musings about mathematical problem-solving research, 1970–1994. *Journal for Research in Mathematics Education*, 25(6), 660–675.

- McKnight, C., Crosswhite, J., Dossey, J., Kifer, E., Swafford, J., Travers, K., & Cooney, T. (1987). *The underachieving curriculum: Assessing U.S. mathematics from an international perspective*. Champaign, IL: Stipes Publishing.
- McKnight, C., Travers, K., & Dossey, J. (1985). Twelfth-grade mathematics in U.S. high schools: A report from the Second International Mathematics Study. *Mathematics Teacher*, 78(4), 292–300.
- Nathan, M. J., Long, S. D., & Alibali, M. W. (2002). The symbol precedence view of mathematical development: A corpus analysis of the rhetorical structure of textbooks. *Discourse Processes*, 33(1), 1–21.
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing office.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication #00-4769). Available at <http://www.nichd.nih.gov/publications/nrp/smallbook.htm>
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.
- National Research Council. (2005). *On evaluating curricular effectiveness: Judging the quality of K–12 mathematics evaluations*. Washington, DC: National Academy Press.
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425-2094, 2002.
- Ridgway, J., Crust, R., Burkhardt, H., Wilcox, S., Fisher, L., & Foster, D. (2000). *MARS report on the 2000 tests*. Palo Alto, CA: Silicon Valley Mathematics Assessment Collaborative.
- Schmidt, W. H., McKnight, C. C., & Raizen, S. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, NL: Kluwer; Washington, DC: National Center for Improving Science Education/WestEd.
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press.
- Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York: Macmillan.
- Schoenfeld, A. H. (2001). Mathematics education in the 20th century. In L. Corno (Ed.), *Education across a century: The centennial volume* (100th yearbook of the National Society for the Study of Education, pp. 239–278). Chicago: National Society for the Study of Education.
- Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher*, 31(1), 3–15.
- Schoenfeld, A. H. (2004). The math wars. *Educational Policy*, 18(1), 253–286.
- Schoenfeld, A. H. (in press). Method. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning*. New York: Macmillan.
- Senk, S., & Thompson, D. (Eds.). (2003). *Standards-oriented school mathematics curricula: What does the research say about student outcomes?* Mahwah, NJ: Lawrence Erlbaum.
- Viadero, D. (2004, August 11). Researchers question Clearinghouse choices. *Education Week*, 30–32.
- What Works Clearinghouse. (2004a). Curriculum-based interventions for improving K–12 mathematics achievement—middle school. Retrieved February 27, 2005, from <http://www.whatworks.ed.gov/>
- What Works Clearinghouse. (2004b). Detailed study report: Baker, J. J. (1997). *Effects of a generative instructional design strategy on learning mathematics and on attitudes towards achievement*. Unpublished doctoral dissertation, University of Minnesota. Retrieved February 27, 2005, from <http://www.whatworks.ed.gov/>
- What Works Clearinghouse. (2004c). Detailed study report: Kerstyn, C. (2001). *Evaluation of the I CAN LEARN Mathematics Classroom: First year of implementation (2000–2001 school year)*. Unpublished manuscript. Retrieved February 27, 2005, from <http://www.whatworks.ed.gov/>
- Whipple, G. M. (Ed.). (1930). *Report of the Society's Committee on Arithmetic* (29th yearbook of the National Society for the Study of Education). Bloomington, IL: Public School Publishing Company.

#### AUTHOR

ALAN H. SCHOENFELD is the Elizabeth and Edward Conner Professor of Education in the Graduate School of Education, University of California, Berkeley, Tolman Hall #1670, Berkeley, CA 94720-1670; e-mail [alans@berkeley.edu](mailto:alans@berkeley.edu). His interests include the study of mathematical thinking, teaching, and learning; research that has an impact on practice; and research methods.

Manuscript received April 1, 2005

Revision received July 2, 2005

Accepted August 7, 2005