Do Randomized Controlled Trials Meet the "Gold Standard"?

A Study of the Usefulness of RCTs in the What Works Clearinghouse

ALAN GINSBURG

AND MARSHALL S. SMITH

March 2016

AEI

AMERICAN ENTERPRISE INSTITUTE

Table of Contents

Executive Summary

The What Works Clearinghouse (WWC), which resides in the Institute of Education Sciences (IES), identifies studies that provide credible and reliable evidence of the effectiveness of a given intervention. The WWC gives its highest rating of confidence to only well-implemented Randomized Control Trial (RCT) designs. RCTs are clearly the "gold standard" to minimize bias in outcomes from differences in unmeasured characteristics between treatment and comparison populations. Yet when the treatment is a complex intervention, such as the implementation of an education curriculum, there is a high potential for other sources of serious estimation bias.

Our analysis of the usefulness of each of the 27 RCT mathematics studies (grades 1–12) meeting minimum WWC standards identifies 12 nonselection bias threats, many of which were identified in a 2004 National Research Council (NRC) report. These nonselection bias threats are not neutralized by randomization of students between the intervention and comparison groups, and when present, studies yield unreliable and biased outcomes inconsistent with the "gold standard" designation.

Threats to the usefulness of RCTs include:

- **Developer associated.** In 12 of the 27 RCT studies (44 percent), the authors had an association with the curriculum's developer.

- **Curriculum intervention not well-implemented.** In 23 of 27 studies (85 percent), implementation fidelity was threatened because the RCT occurred in the first year of curriculum implementation. The NRC study warns that it may take up to three years to implement a substantially different curricular change.

- **Unknown comparison curricula.** In 15 of 27 studies (56 percent), the comparison curricula are either never identified or outcomes are reported for a combined two or more comparison curricula. Without understanding the comparison's characteristics, we cannot interpret the intervention's effectiveness.

- **Instructional time greater for treatment than for control group.** In eight of nine studies for which the total time of the intervention was available, the treatment time differed substantially from that for the comparison group. In these studies we cannot separate the effects of the intervention curriculum from the effects of the differences in the time spent by the treatment and control groups.

- **Limited grade coverage.** In 19 of 20 studies, a curriculum covering two or more grades does not have a longitudinal cohort and cannot measure cumulative effects across grades.

- **Assessment favors content of the treatment.** In 5 of 27 studies (19 percent), the assessment was designed by the curricula developer and likely is aligned in favor of the treatment.

- **Outdated curricula.** In 19 of 27 studies (70 percent), the RCTs were carried out on outdated curricula.

Moreover, the magnitude of the error generated by even a single threat is frequently greater than the average effect size of an RCT treatment.

Overall, the data show that 26 of the 27 RCTs in the WWC have multiple serious threats to their usefulness.

One RCT has only a single threat, but we consider it serious. We conclude that none of the RCTs provides sufficiently useful information for consumers wishing to make informed judgments about which mathematics curriculum to purchase.

As a result of our findings, we make five recommendations. Note that all reports stemming from the five recommendations should be made public.

Recommendation 1: IES should review our analyses of the 27 mathematics curriculum RCTs and remove those that, in its view, do not provide useful information for WWC users. The IES should make their judgments and rationale public.

Recommendation 2: The IES should examine the other curriculum studies and curriculum RCTs in the WWC. The review should be based on the same criteria as in recommendation 1, and the IES should remove those studies that, in their view, do not provide useful information.

Recommendation 3: The IES should review a representative sample of all the other noncurricula RCT intervention studies in the WWC. The review should use the same criteria and standards as in recommendations 1 and 2. Studies that do not meet the standards established for the reviews of the curriculum studies should be removed from the WWC.

Recommendation 4: Evaluations of education materials and practices should be improved. First, the IES should create an internal expert panel of evaluators, curriculum experts, and users (for example, teachers and administrators) to consider how, in the short term, to improve the current WWC criteria and standards for reviewing RCTs in education.

Second, the IES and the Office of Management and Budget (OMB) should support an ongoing, five-year panel of experts at the NRC or the National Academy of Education to consider what would be an effective evaluation and improvement system for educational materials and practices for the future. It should also consider how this system might be developed and supported and what the appropriate role of the federal government should be in designing, creating, and administering this system.

Recommendation 5: OMB should support a three-year study by a panel of unbiased experts and users convened by the NRC to look at the quality of RCT studies in noneducation sectors. We see no reason to expect that RCTs funded out of the Labor Department, HUD, Human Services, Transportation, or USAID would be immune from many of the flaws we find in the mathematics curriculum RCTs in the WWC.

Part I:
Introduction and Context

The What Works Clearinghouse (WWC), instituted in 2002 as part of the Institute of Education Sciences (IES) within the US Department of Education, describes its mission as thus: "The goal of the WWC is to be a resource for informed education decision-making. To reach this goal, the WWC reports on studies that provide credible and reliable evidence of the effectiveness of a given practice, program, or policy (referred to as 'interventions')."[1]

The purpose of our review is to determine how useful randomized controlled trials (RCTs) in the WWC might be in helping teachers and school administrators make accurate, informed decisions about their choice of mathematics curricula. The WWC compiles high-quality evidence on curricula's effectiveness and makes it available online, but for that evidence to be useful, it must present an accurate picture of each curriculum's effectiveness.[2] To analyze this, we examine all intervention studies of mathematics curricula in elementary, middle, and high schools using RCT methodology that were reported on the WWC website on December 1, 2014.

We have no quarrel with the powerful logic and overall potential of the methodology of RCTs. A well-implemented RCT is an important tool for finding an unbiased estimate of the causal effect of a deliberate intervention.

RCTs have been held in high esteem and actively used since the 1920s, when R. A. Fischer used controlled experiments to improve farming, as well as today, as the National Institutes of Health uses RCTs to evaluate the effectiveness of drugs and medical procedures.[3] Experimental psychologists and scientists also make good use of RCTs, which work especially well in highly controlled settings where the character of the intervention and the control groups are very clear. Over the past two decades, many organizations and US government officials have touted RCTs as the best way to produce serious evidence of the effects of various social and educational interventions, labeling RCTs as the "gold standard" for evaluating government programs.[4]

Yet not every statistician and scholar has unilateral faith in RCTs. The first concern is that while a single well-done RCT has internal validity, it is carried out with a particular intervention, for a particular sample, at a particular time, and in a particular place, and it produces a valid estimate of the intervention's effect only in that setting. Therefore, most single RCTs do not have external validity.[5]

To address this issue, William Shadish, Thomas Cook, and Donald Campbell propose that a meta-analysis of multiple trials could help establish external validity.[6] Others argue that the strength of the instructional design and learning theory embedded in the intervention can help guide which RCTs are carried out to establish external validity.[7] But no one argues that the results of a single RCT will necessarily generalize to different populations at different times and places.

Second, a single RCT establishes only one data point within the distribution of estimates of a true "effect" size. The single data point may be atypical, even in the most carefully designed studies. Advocates and skeptics of RCTs urge replication of RCT studies, as Donald Campbell commented in 1969: "Too many social scientists expect single experiments to settle issues once and for all. This may be a mistaken generalization from the history of great crucial experiments. In actuality the significant experiments in the physical sciences are replicated thousands of times."[8]

A National Research Council (NRC) report of curricular effectiveness recommends strongly that RCTs should have at least one replication before the

evidence from the original RCT is used.[9] This caution is almost always followed in the health science field but is all too often ignored in education and other social science areas.[10]

Third, ensuring that an RCT study has internal validity requires more than randomization, appropriate statistical methodology, and replication. Many scholars have identified problems with inferences drawn from RCTs that are used to determine the effectiveness of interventions in settings sensitive to a wide variety of design and implementation threats. Again returning to Campbell: "We social scientists have less ability to achieve 'experimental isolation,' because we have good reasons to expect our treatment effects to interact significantly with a wide variety of social factors many of which we have not yet mapped."[11]

Implementing a curriculum is complex. Effectively using a curriculum requires deep understanding of its content and pedogogy and of the various instructional needs of the 20 to 30 different students in a classroom. Moreover, a school or classroom environment is complex and subject to minor and major disruptions. These characteristics produce special evaluation challenges.

Complex interventions have interacting components within the intervention and the environment and require behaviors that are difficult to implement effectively. Because of these components, good experimental design requires rigorous attention to numerous internal validity threats to the study's usefulness. The British Medical Research Council's evaluation recommendations for complex interventions include: "A good theoretical understanding is needed of how the intervention causes change" and that "lack of effect may reflect implementation failure (or teething problems) rather than genuine ineffectiveness; a thorough process evaluation is needed to identify implementation problems."[12]

This report examines 12 potential threats to the usefulness of the 27 RCT mathematics curriculum studies (grades 1–12) that were in the WWC on December 1, 2014. From our examinations of possible threats, we ask whether the RCTs offer credible and reliable evidence and useful knowledge for making decisions about state, school, or classroom curricula.[13] We conclude with some general observations and five recommendations to the IES and the Office of Management and Budget (OMB) in the federal government.

The Context

This report uses data from the What Works Clearinghouse, a website of the US Department of Education.[14] We have also gone to original sources to better understand the original studies or obtain additional information about the authors' association with a curriculum's publishers. We examine individual RCTs rather than the WWC itself, although we reference the WWC many times.

Others have also evaluated the WWC and its use of RCTs. In 2004, in response to a congressional inquiry, the NRC, an arm of the National Academy of Sciences, named a panel of scholars to analyze and report on the effectiveness of mathematics curricula funded by the National Science Foundation.[15] In a separate study, Jere Confrey, the chair of the NRC report, compared the NRC report criteria to the WWC evaluation criteria for individual studies.[16] Confrey concluded that the WWC evaluation criteria were far too narrow when compared with the comprehensive criteria that the NRC recommends.

Alan Shoenfeld, a reviewer of the NRC study, also analyzed the WWC and concluded that "methodological problems rendered some mathematics reports potentially misleading and/or uninterpretable."[17] In 2013, Alan Cheung and Robert Slavin published an article that reviewed the use and effectiveness of technology-based curricula in mathematics.[18] All these articles are relevant and have some overlap with this review.

It has been more than a decade since the NRC report and the Confrey review of the WWC, and the WWC is almost 15 years old, which is long enough to work through initial design and implementation problems. It seemed a fair time for us to look at one aspect of the effort, but we did not try to examine the entire WWC. We looked at the information from only the 27 RCT intervention studies of elementary, middle, and high school mathematics curricula, and we specifically

Table 1. Mathematics Curricula That Meet Review Standards or Meet Standards with Reservations

Math Curriculum	Meet WWC Standards Without Reservations	Meet WWC Standards with Reservations
Elementary		
Odyssey Math		1
Accelerated Math	1	
enVisionMATH	1	
DreamBox Learning	1	
Progress in Mathematics @ 2006	1	
Saxon Math	1	
Investigations in Number, Data, and Space	1	1
Peer-Assisted Learning Strategies	1	
Scott Foresman-Addison Wesley Elementary Mathematics	3	
Middle and High School		
The Expert Mathematician	1	
Saxon Math	1	1
I CAN Learn Pre-Algebra and Algebra	1	2
Accelerated Math		1
Transition Mathematics	1	
PLATO Achive Now	1	
University of Chicago School Mathematics Project (UCSMP) Algebra		1
Cognitive Tutor Algebra	3	1
Cognitive Tutor Geometry	1	
Total	**19**	**8**

Note: Table 1 combines all Cognitive Tutor math RCTs for algebra as a group and separates the Cognitive Tutor for geometry. By contrast, the WWC separates the algebra Cognitive Tutor that serves junior high school from the others, which serve high school. However, all algebra programs served grade nine students, including the junior high school intervention, so we combined them into one Cognitive Tutor category. Unlike the WWC, which grouped the Cognitive Tutor algebra and geometry interventions together, we considered these as different interventions and treat them as separate categories.
Source: What Works Clearinghouse intervention reports, http://ies.ed.gov/ncee/wwc/findwhatworks.aspx.

focused on potential threats to the quality of the RCT estimates of the 27 interventions' effectiveness.[19]

In an RCT intervention study, an RCT treatment outcome is measured against a comparison outcome, and the RCT fully meets the WWC's standards of internal validity or meets the standards with reservation. Three RCTs contained more than one RCT intervention curriculum, and in these cases, we count each comparison as a separate RCT study.[20] Thus there are only 23 separate RCTs and 27 RCT curriculum studies.

We are satisfied with the studies' statistical analyses as corrected by the WWC statisticians. The WWC also checks RCTs for baseline equivalence of treatment and comparison groups and for excessive attrition of subjects during the experiment. The evaluators or the WWC make any needed additional statistical tests and

adjustments to the analyses where either baseline differences or attrition appears excessive. The WWC also rigorously requires adjustments of statistical tests for clustering of students within schools and classrooms.[21]

There are 18 different curricula. Some curricula were involved in more than one RCT. Table 1 names the curricula and the number of RCT studies in two different categories: 19 studies have RCTs that meet all WWC standards for quality of randomization and statistical procedures without reservations, and 8 studies do not fully meet all standards but are deemed by the WWC as usable with reservations.[22]

Methodology

We gathered all the available material from the WWC website on each RCT study, as well as online original reports or relevant articles.[23] In a few cases, we emailed the authors to ask them to clarify something related to their study.

We specifically examined each RCT study against certain criteria to gauge whether it had threats to the standard of credible, reliable, and useful evidence.[24] For all the RCTs, we were able to get a reliable and arguably accurate reading on all or many of the criteria.

Table 2 is a sketch of our conceptualization of the desired criteria for a useful RCT. We have eight general categories where threats may occur: strong theory, independence of association with curriculum developer, curriculum well-implemented, identified comparison group, appropriate grade coverage, objective measurement of outcomes, curriculum not out of date, and replication.

This report does not explore whether the 27 RCT studies met the six crosshatched criteria in Table 2. We assume that the WWC staff in the IES fully reviewed and accepted three traditional statistical criteria: unbiased sampling procedure, sample attrition, and appropriate statistical analyses.

In addition, we did not explore whether the intervention was based on a reasonable theory, if the implementation had been replicated, or whether it used a single or double blind structure, even though these criteria are important.[25] None of the WWC summaries provided adequate information about any of these three potential threats. A few of the research articles discussed the theory of the intervention, but this was atypical and did not provide enough detail to review.

This leaves us with six of the eight categories and 12 potentially threatened criteria. They are:

Study independent of association with curriculum developer

1. Evaluators are independent of association with curriculum developer.

Curriculum well-implemented

2. Curriculum is not studied in first year of implementation.
3. Available evidence of implementation shows fidelity.

Comparison identified

4. Outcomes are reported by each identified curriculum.
5. Comparison and intervention have equivalent instructional time.

Appropriate grade coverage

6. Study is longitudinal, if appropriate.
7. There is broad coverage of a multigrade curriculum.
8. WWC reports outcomes by grade, where available in the original study.

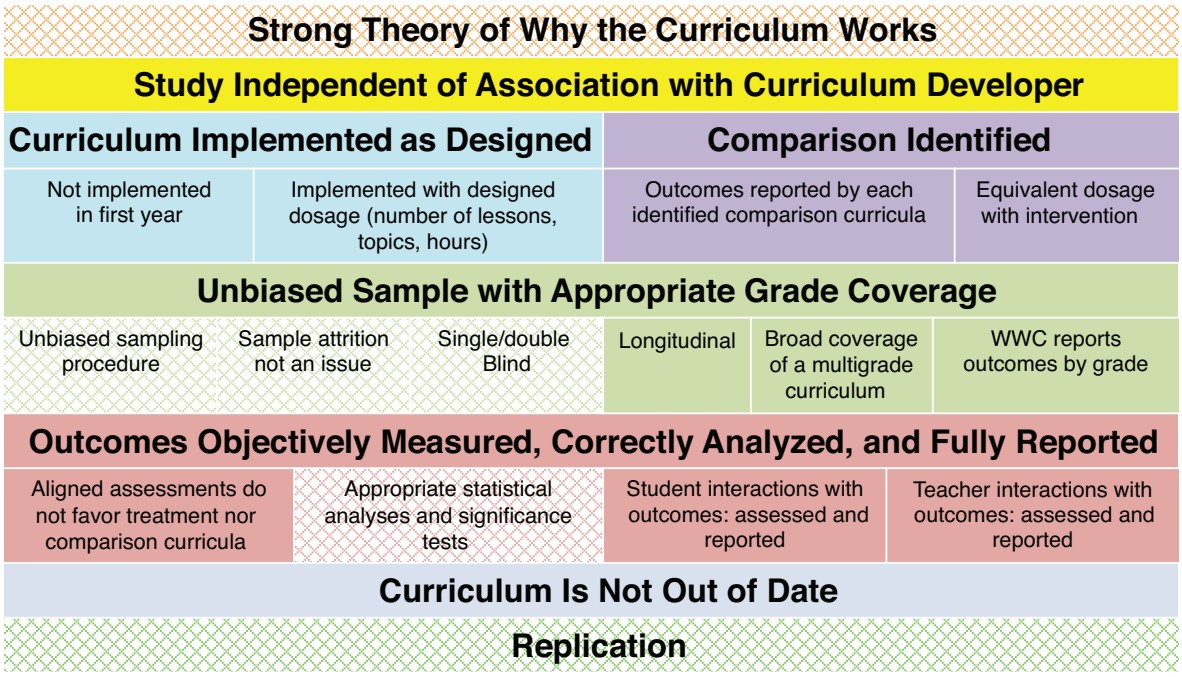
Outcomes objectively measured, correctly analyzed, and fully reported

9. Aligned assessments do not favor treatment or comparison curricula.
10. Student interactions with outcomes are assessed and reported, where available.
11. Teacher interactions with outcomes are assessed and reported, where available.

Curriculum is not out of date

12. Curriculum is current: it is still published, was initially released after 2000, and has not been replaced by a career- and college-ready state standards version (such as the Common Core).

Table 2. Criteria for a Useful RCT Curriculum Study



Note: Cross-hatchers are criteria not examined in this study, because IES reviews adequately addressed these criteria or we did not have data to examine them.
Source: Authors.

When these criteria are not met, threats to the RCT may develop. We describe all 12 of these threat criteria in the following results section, and we evaluate whether they pose threats to the usefulness of any of the 27 RCT studies.

Part II:
Threats to the Usefulness of RCTs

Establishing the extent of a relationship between a curriculum's use and its outcomes scientifically requires ruling out other plausible explanations for this relationship. Russ Whitehurst, who as director of IES actively supported the IES and WWC's focus on RCTs, gave the following rationale in his 2003 speech before the American Education Research Association:

> Randomized trials are the only sure method for determining the effectiveness of education programs and practices. We now have compelling evidence that other methods can lead to estimates of effects that vary significantly from those that would be obtained from randomized trials, nearly always in size and sometimes in direction of effect.[26]

This is a traditional justification for using RCTs. The logic of RCTs arising from the use of random samples provides his evidence, but this ignores the reality that complex social systems such as classrooms and schools often create major challenges for applying RCTs.

Hence, our focus is on determining whether RCT mathematics studies that meet the WWC standards contain one or more of the 12 nonselection bias threats. These nonselection bias threats fall outside of the student randomization process. Consequently, these threats are not neutralized by randomizing student placement in the intervention and comparison groups. If such a threat exists, we cannot say that the intervention produced a reliable or valid effect-size difference in outcomes between the intervention and control groups (that is, the gold standard).

We begin our discussion of each threat criterion by analyzing its description where applicable in the NRC report or the *WWC Procedures and Standards Handbook*.[27] We then summarize the threat's prevalence within each of the 27 RCT curriculum studies included in the WWC mathematics reports in December 2014. The discussion concludes by demonstrating how the failure to address that potential threat would diminish confidence of WWC customers in the accuracy or usefulness of the RCT findings. Where quantitative information is available, usually for a subset of studies affected by the threat, we try to give a sense of the potential magnitude of each threat. We do not consider how interactions of the threats might constitute new additional threats.

A note of caution: our quantification of a potential threat criterion's frequency required, in many cases, a search of the original study documents. Only 18 of these original documents were available, so our estimates present a minimum prevalence count of potential methodological weaknesses for WWC reported studies. We assumed *no threat* unless we had a clear reason to believe that the criterion was violated.

Finally, we have treated two sources of threats together. One source arises directly from RCT design and implementation, with eight threats in this category. The other type of threat arises from limitations in the reporting of the RCT findings by the WWC, with four threats in this category.[28]

The following sections present our findings about the presence of the 12 criteria (threats), organized into the six categories we examined.

Evaluators Associated with the Curriculum Developer

One rule for summative evaluations is that the evaluator should not have a personal stake in the evaluation's outcomes. In studying commercially produced

Table 3. RCT Mathematics Studies, by Association with the Curriculum's Developer or Publisher

Intervention	Yes, an Association (Number of Studies)	No Association (Number of Studies)
Odyssey Math (Elementary)		1
Accelerated Math (Elementary)	1	
enVision Math (Elementary)	1	
DreamBox Learning (Elementary)		1
Progress in Mathematics @ 2006		1
Saxon Math (Elementary)		1
Investigations (Elementary)	1	1
Peer Assisted Learning Strategies (Elementary)	1	
SFAW (Elementary)	2	1
The Expert Mathematician (Middle)		1
Saxon Math (Middle)	1	1
I CAN Learn (Middle)	3	
Accelerated Math (Middle)	1	
Transition Math (Middle)		1
Plato (Middle)		1
UCSMP Algebra		1
Cognitive Tutor Algebra	1	3
Cognitive Tutor Geometry		1
Total	**12**	**15**

Note: Association includes the study being sponsored or funded by a curriculum developer, one or more study authors being currently or formerly employed by the curriculum developer, or the curriculum developer having a major role in reviewing study content.
Source: Drawn from What Works Clearinghouse intervention reports or original full-study reports.

curricular evaluations, a particular risk is curriculum developers who exert strong influence over the study and have personal or financial interest in showing their curriculum favorably.

The NRC report warns of this:

> The relationship of an evaluator to a curriculum's program designers and implementers needs to be close enough to understand their goals and challenges, but sufficiently independent to ensure fairness and objectivity. During stages of formative assessment, close ties can facilitate rapid adjustments and modifications to the materials. However, as one reaches the stage of summative evaluation, there are clear concerns about bias when an evaluator is too closely affiliated with the design team.[29]

A close evaluator-developer relationship can bias RCT studies toward overly favorable estimates without blatant number altering. An evaluator might randomly assign students to teachers, but consciously or unconsciously select stronger teachers for the intervention and weaker teachers for the comparison.[30] A favored treatment might receive extra dosage, resources, and professional training. Outcome assessments may also overly align with the content of the treatment intervention while not aligning as well with the content taught to the comparison group.

Table 4. Effect-Size Estimates, by Whether RCT Mathematics Studies Are Associated with Curriculum Developer

Intervention	Yes, an Association	No Association
Investigations (Elementary)	0.12	–0.04
SFAW (Elementary)	0.01	–0.09
Saxon Math (Middle)	0.19	0.14
Cognitive Tutor Algebra	0.38	–0.14
Average	**0.18**	**–0.03**

Source: What Works Clearinghouse intervention reports or original full-study reports.

The file-drawer option is an additional possibility, in which the authors or those contracting for the study never publish results perceived as unfavorable. This may also happen because publishers favor studies with positive findings. In terms of the quantitative importance of publication bias, one analysis of social science research presented in *Science* found that in a group of 221 NSF-supported studies, those with "strong results are 40 percentage points more likely to be published than are null results."[31]

When there is concern over lack of study independence, replication could serve a particularly crucial validation role, provided the replication is conducted with investigators who are independent of the prior investigation and developer.[32] However, independent study replication is not an explicit criterion in WWC reviews.

We reviewed the IES *WWC Procedures and Standards Handbook* and the *Reporting Guide for Study Authors,* and neither discusses the issue of whether the evaluation-study authors are independent of the curriculum-intervention developers.[33] Consequently, WWC intervention reports do not identify whether a study has an association with a developer.

To their credit, several authors do independently disclose such a relationship in their reports.[34] When direct information was not available, we inferred authorship from context, including the authors' website work history or whether the authors were with an evaluation company and likely to do studies under contract support from the curriculum developers.[35]

Table 3 shows that the potential for bias from developer association is widespread. Nearly half of the RCT curriculum studies in mathematics are classified as developer associated.

Four WWC mathematics curricula were examined by both developer-associated and developer-independent studies. For these four, the effect-size estimates from the developer-associated studies can be compared with those from independent studies (Table 4). In all four cases, the effect-size estimates are more positive for the developer-associated estimates. The average quantitative advantage of developer-associated studies is about two-tenths of a standard deviation, which for most studies is as large as or larger than the effect-size advantage of the treatment.[36]

The Curriculum Intervention Is Not Well-Implemented

Another threat to obtaining credible and reliable evidence—and hence to the usefulness of the WWC RCT findings—is evaluating a poorly implemented math intervention. The NRC report on evaluating the effectiveness of mathematics curricula recommends that "evaluations should present evidence that provides reliable and valid indicators of the extent, quality, and type of the implementation of the materials."[37] The NRC report indicates that fidelity of implementation is commonly measured by "the extent of coverage of the curricular material, the consistency of the instructional approach to content in relation to the program's theory, reports of pedagogical techniques, and the length of use of the curricula."[38]

The *WWC Reporting Guide for Authors* requires a discussion on implementation as part of a study's submission: "Describe the actual implementation of the intervention studied, including adaptations of content, level and variation in duration and intensity."[39] However, in practice, the information on implementation that the WWC reports is scattershot, and implementation fidelity is not considered in the development of effectiveness ratings.

This section looks at two of the NRC indicators of implementation fidelity chosen as proxies for implementation quality because they were most readily available from the WWC intervention reports or the full-study original reports.

Treatment Implemented in First Year of Study. In their classic evaluation textbook, Peter Rossi, Howard Freeman, and Mark Lipsey offer guidance for assessing the impact of an intervention: "interventions should be evaluated for impact *only when they have been in place long enough to have ironed out implementation problems*."[40] (Italics added.)

Similarly, the NRC suggests that "for curricula that are quite discontinuous with traditional practice, particular care must be taken to ensure that adequate commitment and capacity exists for successful implementation and change. *It can easily take up to three years for a dramatic curricular change to be reliably implemented in schools*."[41] (Italics added.)

The WWC reporting guide does not request information about whether teachers had prior experience using the treatment curriculum for one or more years before assessing student outcomes. However, the WWC reports usually present information about the year the random assignment started and when the curriculum treatment launched, so it is possible to accurately infer whether the treatment curriculum was evaluated in its first year of use.

Analyses of the intervention reports indicate that in 23 of the 27 RCT studies in mathematics (85 percent), the treatment is assessed in the teacher's first year of implementation. That is, the study process does not build in any prior use of the treatment curriculum by teachers. Few study authors addressed the issue, but some did note challenges of first-year implementation.

For example, Pane et al. discusses the difficulty that teachers without prior experience had in implementing blended learning to support the Cognitive Tutor's learner-centered instruction:

> Researchers also collected observation and interview data on teachers' instructional practices. These data suggest that many teachers had difficulty implementing the treatment curriculum's learner-centered pedagogy. In fact, observed levels of learner-centered practices were only modestly higher in treatment classes than in control classes. In both treatment and control classes, however, higher levels of learner-centered pedagogy were associated with higher student achievement.[42]

Only four RCTs involved schools that explicitly had prior exposure to the treatment (Table 5). Prior implementation occurred through two types of RCT assignment strategies.

Odyssey Math and Odyssey Reading were supplemental instructional programs already in place in the evaluation-participating schools before the start of the evaluation.[43] During the summative evaluation, students were randomly assigned to receive either the supplemental Odyssey Math or Odyssey Reading program, but not both. The students also continued to receive the common regular math and reading programs taught in the schools.

Thus, the Odyssey Reading group served as the control for the math treatment group, which used Odyssey Math, and both were in place before the launch of the experimental evaluation.[44] For many students, this comparison was in essence comparing an extra year of Odyssey Math for the treatment, when both the treatment and comparison students have already had a year or more of prior exposure to Odyssey Math.

The other three studies employed different cohorts of students over two or more years covering the same grades. The Plato and the Cognitive Tutor Algebra studies (Table 5) were for two years, with the first year providing teachers with a year of experience for the second year. The Cognitive Tutor Geometry study had a mix of schools by length of participation. Those with only one year of participation had no prior implementation experience; the schools with

Table 5. RCT Mathematics Studies Implemented with One or More Years of Prior Treatment Use

Intervention	Authors	Description of Prior Use
Odyssey Math	DiLeo	Odyssey Math and Reading, a supplemental technology program, was in place in schools for one or more years prior to the RCT. The treatment used Odyssey Math and the comparison used Odyssey Reading during the supplemental instruction period.
Plato	Campuzano, Dynarski, Agodini, and Rall	Partial prior use of treatment for two cohorts at same grade: first-year treatment was implemented without prior use; second-year implementation was after one year of use.
Cognitive Tutor Algebra	Campuzano, Dynarski, Agodini, and Rall	Partial prior use of treatment for two cohorts: first-year treatment was implemented without prior use; second-year implementation was after one year of use.
Cognitive Tutor Geometry	Pane, McCaffrey, Slaughter, Steele, and Ikemoto	Partial prior use of treatment for five of eight high schools: two high schools participated in each of three academic years with two years of prior use of Cognitive Tutor; three high schools participated for two years with one year of prior use of Cognitive Tutor; and three high schools participated for only one year with no prior use built in to study.

Source: What Works Clearinghouse intervention reports; Judy DiLeo, "A Study of a Specific Language Arts and Mathematics Software Program: Is There a Correlation Between Usage Levels and Achievement?" (doctoral dissertation, Indiana University of Pennsylvania, May 2007); Larissa Campuzano et al., *Effectiveness of Reading and Mathematics Software Products: Findings from Two Student Cohorts,* National Center for Education Evaluation and Regional Assistance, February 2009; and John F. Pane et al., "An Experiment to Evaluate the Efficacy of Cognitive Tutor Geometry," *Journal of Research on Educational Effectiveness* 3, no. 3 (January 2010): 254–81.

two years had one year of experience with the Cognitive Tutor; and the schools with three years offered two years of experience.

Prior use of the treatment curriculum may also occur because of teachers' chance experience with the curriculum. One study (Table 6) surveyed teachers and found that the percentage of teachers who previously used their randomly assigned treatment curriculum ranged from only 4 percent for Math Expressions to 21 percent for Scott Foresman–Addison Wesley (SFAW), a widely used mathematics curriculum.[45] Based on the teacher-experience differential, the SFAW and Saxon curricula had an advantage, because a greater percentage of their teachers had prior exposure to the curriculum.

These examples suggest that it is occasionally feasible to conduct RCT studies where schools have at least one year of prior intervention experience. It is usually not done, at least in part because it typically requires an expensive multiyear experiment, large student mobility, and teacher assignments to the intervention that are sustained over several years.

Implementation Does Not Show Fidelity. When evaluators in the WWC studies measure implementation fidelity of a treatment curriculum, they typically measure actual to intended dosage. Table A1 highlights the studies we identified as directly measuring implementation fidelity, the measures used, and key findings.

Table 6. Percentage of Teachers Who Had Previously Used Treament Curriculum

	All Teachers	Investigations	Math Expressions	Saxon	SFAW	p-value
Used the Assigned Curriculum at the K–3 Level Before the Study	11.5	5.5	3.6	16.2	21.1	0.01

Source: Roberto Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders,* National Center for Education Evaluation and Regional Assistance, October 2010.

Fourteen intervention studies—more than half—included some form of implementation-fidelity measure. Of those 14 studies, 13 adopted implementation-fidelity measures that incorporate a dosage indicator (for example, minutes of actual instruction compared to intended minutes, percentage of intended instructional components used over some regular time period, or percentage of intended objectives or lessons completed). Data from 8 of the 14 RCT curriculum studies indicate low implementation fidelity based on each study's measures in Table A1. Six of these studies also found a relationship between the degree of implementation fidelity in the intervention classrooms and student outcomes.

A typical example is the study of enVisionmath, which employed a two-component composite indicator of implementation: completion of key program components and the percentage of enVisionmath topics completed.[46] The full RCT report of enVisionmath observed that "students whose teachers implemented the major components of enVisionmath with high fidelity showed greater improvement than students of teachers who implemented envision Math with low fidelity."[47]

Accelerated Math is a second example of a serious implementation issue, which could have been a source of the curriculum's null effects. In this study, 40 percent of the students in the Accelerated Math experimental group never received the program. Among students who received Accelerated Math, those exposed to high implementation, as measured by the number of objectives mastered, had greater achievement gains than students exposed to low implementation. The WWC reported only the average achievement gain for all students.

Unlike the prior two studies, the RCT study evaluating I CAN Learn has *extreme* implementation weaknesses.[48] This RCT met WWC criteria without reservations, but its lesson completion rate (a common measure of implementation fidelity) is so low that it is surprising that the study produced any significant positive results.

I CAN Learn students began using the yearlong curriculum in only the second semester. Additionally, they are expected to complete about 100 lessons during the school year, but study students completed only 12.1 lessons on average. Yet the WWC approved the study in which I CAN Learn students outperformed the comparisons by a rather large 0.35 effect size. We suspect that this RCT design was flawed in other ways to produce the positive I CAN Learn results.[49]

Unclear Comparison Curricula

WWC impacts are measured as the difference in outcomes between the treatment and comparison (counterfactual) curricula. Hence, in estimating treatment effectiveness, a key question is what the comparison curriculum was against which effectiveness is measured. Was it a high- or low-quality curriculum? Was it conceptually different than the intervention curriculum? Was it a mixture of curricula?

This choice is crucial to the outcome of the RCT comparison. For example, student outcomes from a basic curriculum plus a supplemental technology program can be compared against a counterfactual with another supplemental technology program, or they can be compared to a counterfactual program without technology. The comparisons are likely to yield different effectiveness estimates.[50]

This section explores two common threats to evaluations that emerge from the nature of the comparison intervention in randomized experiments: (1) the

comparison curricula are not identified, or if identified, outcomes are not reported for each comparison; and (2) the treatment and comparison instructional time are not equivalent in creating a fair playing field for the comparison intervention. This is different from the implementation-fidelity threat, in which the actual treatment dosage was viewed as adequate or inadequate relative to the intended dosage. Here, the criterion is with respect to the comparability of the treatment curriculum with the comparison dosage.

Unidentified Comparison Curricula. The NRC report stated its concern with identifying the comparison curricula as follows:

> We express concern that when a specified curriculum is compared to an unspecified content which is a set of many informal curriculum, the comparison may favor the coherency and consistency of the single curricula, and we consider this possibility subsequently under alternative hypotheses. We believe that a quality study should at least report the array of curricula that comprise the comparative group and include a measure of the frequency of use of each, but a well-defined alternative is more desirable.[51]

The *WWC Standards and Procedures Handbook* discussion of the comparison intervention is quite general and focuses on reporting with multiple comparisons:

> The WWC generally considers any contrast related to the intervention of interest when reviewing a study. For example, a study may have three groups (intervention Y, intervention Z, and a comparison group). A product focused on intervention Y may include only the contrast with the comparison group, or it may also include the contrast with intervention Z. Similarly, although a study may examine the effects of intervention Y relative to a comparison group receiving intervention Z, a WWC review focused on intervention Z would include this study by viewing Z as the intervention condition and Y as the comparison.[52]

Overall, the WWC reports run the gamut in describing the comparison curricula, ranging from no information about the comparison math programs to specifically comparing the treatment intervention against a single, named comparison math program. More precisely, 15 of the 27 studies (56 percent) did not report outcomes against a specific comparison curriculum. Many of these studies do not name the comparison curricula. A few name several curricula but do not report results against any one specific curriculum (Tables A3 and A4).

Agodini et al. offer real-world insights into how the choice of the counterfactual curricula affects effectiveness estimates.[53] This RCT study compares the effectiveness of 12 possible paired combinations of curricula obtained from four different named math interventions for grades one and two. These interventions were chosen to represent different math intervention types.

Table 7 (taken from the original report) displays how the estimated statistical significance of a curriculum's comparative effect size in the Agodini et al. study differs depending on the comparison curriculum. It also shows how statistical significance is influenced by different statistical analyses.

When each comparison is considered as statistically independent, four comparisons meet the 0.05 statistical significance level. Math Expressions is statistically effective in both grades, compared with SFAW; compared with Investigations, only effective at grade one; and never significant, compared with Saxon Math. Saxon Math is statistically effective for only one of six comparisons: at grade two compared with SFAW. When the statistical results are adjusted for the six multiple comparisons made at each grade level, only the Saxon-SFAW differential of 0.17 standard deviations at second grade is significant.[54]

Differences in Dosages. When dosage levels differ between the treatment and comparison students, it is no longer possible to separate the effect of different dosage amounts on student achievement from that caused by the differing nature of the treatment and counterfactual math curricula.

Interestingly, neither the NRC nor the WWC review criteria discuss differences in instructional dosage as a confounding factor. The NRC report does address additional professional development as "inflating

Table 7. Differences Between Pairs of Curricula in Average HLM-Adjusted Student Achievement (in Effect Sizes) for First and Second Graders

	Effect of					
	Investigations Relative to			Math Expressions Relative to		Saxon Relative to
	Math Expressions	Saxon	SFAW	Saxon	SFAW	SFAW
First-Grade Classrooms (Effect Size)	–0.11*	–0.07	0.00	0.05	0.11*	0.07
Second-Grade Classrooms (Effect Size)	–0.03	–0.09	0.09	-0.05	0.12*	0.17*
Average effect size across all six comparisons: Saxon Math: 0.07 (significant at 0.05); Investigations: –0.04 (not significant at 0.05); SFAW: –0.09 (significant at 0.05); Math Expressions is not available.						

Note: Effect size is significant at the 0.05 level.
Source: Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula*.

treatment," but our review suggests that an even stronger case could be made for additional instructional dosage based on several of the curriculum intervention reports.[55] The WWC reports summarizing intervention characteristics also include a section on "staff/teacher training," and these typically showed some modest advantage for the treatment.[56]

Estimates of instructional time were retrieved from the original study reports and from the WWC write-ups for 9 of the 27 WWC intervention comparisons (Table A2). Five of the interventions for documented instructional time involved technology use with added time between one and a half to three hours per week. Only one study of a supplementary intervention leveled the playing field by reducing the treatment group's regular math instructional program by the amount of time students were exposed to the supplementary treatment.[57]

Agodini et al. compared the effectiveness of four curricula.[58] At the second-grade level, Saxon Math had a dosage time of about 6.9 hours and comparison groups had times of about 5.5 hours, a 25 percent advantage for Saxon Math. In first grade, Saxon Math instructional time was 6.1 hours, compared with about 5.1 hours for the comparison, a 20 percent advantage. This confounds the Saxon positive effect-size comparisons with Saxon's greater instructional time.

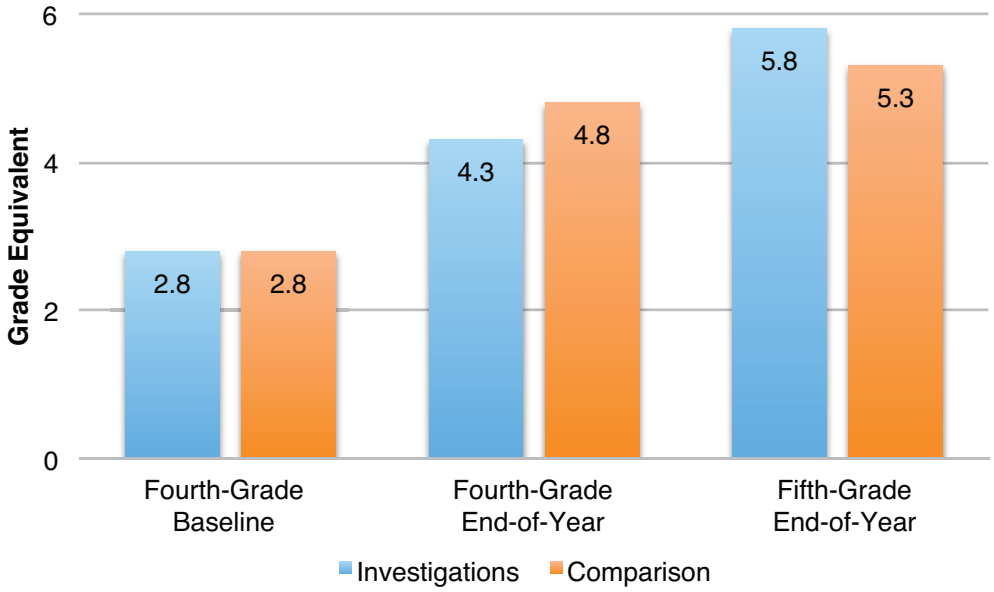
The bottom line is that instructional time matters and that a level, unbiased playing field is one with treatment and comparison interventions approximately equal in dosage time. This was not the case for some of the WWC intervention studies that included measures of instructional time. WWC customers would most likely want to know whether improvements occurred in conjunction with additional time.

Multigrade Curricula Not Adequately Studied

WWC customers would ideally like to know a curriculum for students who have had exposure since the beginning grade of the curriculum. Accordingly, this section explores two study-design threats for the coverage of student outcomes across grades within WWC randomized experiments: (1) the absence of a longitudinal student cohort across grades; and (2) inadequate coverage of a broader multigrade curriculum.

No Longitudinal Cohort. A multiyear longitudinal study tracking the same student cohort through multiple grades is designed to explore the cumulative effects of a curriculum. The NRC report states its advantages: "They improve measurement by being sensitive to

Figure 1. One- and Two-Year Longitudinal Results for Investigations Versus Comparison Curricula, Grades Four and Five



Source: Gatti and Giordano, *Investigations in Number, Data, and Space Efficacy Study.*

accumulation of effects and can indicate whether the rate of learning holds steady or changes with increased instructional time."[59]

The *WWC Procedures and Standards Handbook* omits the merits of longitudinal studies and discusses only the formal procedural aggregating and reporting of outcomes measured at different points in time.[60]

A longitudinal design also has trade-offs; it is expensive and time-consuming and risks losing high percentages of students in both intervention and control groups. Perhaps reflecting the difficulties of conducting multiyear RCTs, among the 20 math studies for interventions covering two or more grades, only one math study reported more than one year of test scores for the same students. This longitudinal study of Investigations analyzed a first- and fourth-grade cohort for two years.[61] The two-year longitudinal study suffered from severe attrition, with only 60 percent of its first-grade students and 72 percent of its fourth-grade students remaining in the sample in its second year. The WWC affirmed these results with reservations.[62]

Figure 1 illustrates the Investigations findings' sensivity to having the second-year results. At the end of the first year, the fourth-grade comparison students outperformed Investigations students in terms of grade equivalence by half a grade. However, by the end of the second year, the differential had reversed itself, and the students using Investigations outperformed the comparison by half a grade equivalent. Interestingly, students have two years of exposure to the curriculum, but teachers at each grade had only one year. The WWC intervention report for Investigations displays the significant and positive second-year findings, but WWC did not report the negative first-year results.

Inadequate Coverage of a Multigrade Curriculum. Compared to a longitudinal approach, a less time-consuming design for tracking students would compare student growth over a single year for some portion or cross section of each grade in a multigrade curriculum package. This design is a set of one-year

Table 8. Grade Coverage of RCT Studies for Elementary Math

Curriculum Intervention	Study	Grades Intervention Covers	Grades Study Covers	Inadequate Coverage of Grades of a Multigrade Curriculum
Odyssey Math	DiLeo	K–5	5	Y
Accelerated Math	Ysseldyke and Bolt	1–5	2–5	N
enVisionmath	Resendez and Azin	1–5	2 and 4	N
DreamBox Learning	Wang and Woodworth	K–5	K and 1	Y
Progress in Mathematics 2006	Beck	K–5	1	Y
Saxon Math	Agodini et al.	K–5	1 and 2	Y
Investigations in Number, Data, and Space	Agodini et al. Gatti and Giordano	K–5 K–5	1 and 2 1, 2, 4, and 5	Y N
Peer-Assisted Learning Strategies	Fuchs et al.	1–5	1	Y
Scott Foresman–Addison Wesley Elementary Mathematics	Agodini et al. Resendez and Azin Resendez and Manley	K–5 K–5 K–5	1 and 2 3 and 5 2 and 4	Y N N

Source: WWC online math intervention reports; DiLeo, "A Study of a Specific Language Arts and Mathematics Software Program"; Ysseldyke and Bolt, "Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement"; Miriam Resendez and Mariam Azin, *A Study on the Effects of Pearson's 2009 enVisionmath Program*; Wang and Woodworth, *Evaluation of Rocketship*; Beck Evaluation & Testing Associates Inc., *Progress in Mathematics 2006: Grade 1 Pre-Post Field Test Evaluation Study,* Sadlier-Oxford Division, William H. Sadlier Inc., 2005; Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula*; Gatti and Giordano, *Investigations in Number, Data, and Space Efficacy Study;* L. S. Fuchs et al., "Enhancing First-Grade Children's Mathematical Development with Peer-Assisted Learning Strategies"; M. Resendez and M. Azin, *2005 Scott Foresman–Addison Wesley Elementary Math Randomized Control Trial: Final Report,* PRES Associates Inc., 2006; and M. Resendez and M. A. Manley, *Final Report: A Study on the Effectiveness of the 2004 Scott Foresman–Addison,* PRES Associates Inc., 2005.

outcome measures on the treatment and comparison samples at different grades.

The grades reported might be all the grades covered by the curriculum package or the front or back end of the curriculum grade range. Of course, the various cohorts preferably would be roughly equivalent. However, if the upper-grade cohorts have not had the particular intervention throughout their school experience, then the end-grade effect size is not a measure of cumulative benefits but of only a single grade change.

Having a broad range of grade coverage is particularly important in elementary school. The early mathematics content (K–2) stresses basic arithmetic, computational fluency, and knowledge and recall of common shapes and measures. In the later elementary grades (3–5), the mathematical concepts and problems become more complex, with a focus on fractions for arithmetic, coordinate planes in geometry, and area and volume for measurement. Thus, a curriculum yielding positive effect sizes in the early grades may not carry these results over into later grades.

Along these lines, it is worth remembering when the federal government invested more than $5 billion on the Reading First program, which was based heavily on RCT reading research and stressed a foundation in early-grade reading skills with phonics-based approaches. However, the final evaluations of Reading First showed that, while the program was indeed successful at changing teacher practices, these changes did not transfer into gains on higher-order reading comprehension—although there was some evidence of improvement in the more basic early decoding skills.[63]

At the elementary grades (K–5), all 12 WWC studies are of a multigrade curriculum for either grades K–5 or 1–5. For each curriculum, Table 8 displays the grades

that the curriculum intervention covers and that the study covers. At the elementary level, we consider adequate coverage as a study with at least one grade sampled between K–2 and between 3–5. Seven of the 12 elementary math studies failed to meet this criterion.

At the middle school grades (6–8), adequate study coverage is defined as covering at least two of the three grades. Six of the 10 middle school curricula cover only a single grade of a three-grade middle school curriculum and fail to meet the review standard (Table A4).

At the high school grades (9–12), programs are typically a single year, and we considered it appropriate to judge the usefulness of a high school algebra or geometry program from a single-grade sample. Hence, all five high school studies met the grade-coverage criterion (Table A4).

Outcomes Not Fully Reported by Grade. While some RCT mathematics studies report only one set of average outcomes across several grades, others have computed results grade by grade. In these cases, the WWC reports are inconsistent. Sometimes the WWC publishes the individual grade outcomes, but frequently, grade-by-grade outcome breakouts for a multigrade curriculum are not published, even when they are available. Thus, the usefulness of WWC reports diminishes, because customers lose potentially important information about effect-size estimates at different grades. Looking across the 27 studies, 6 were identified with outcome estimates for two or more individual grades and for which the WWC publishes only average scores.

An example is the enVisionmath study, which randomly assigned students in the second and fourth grades. The study explored differences among various student subgroups, including by grade level. "While 4th grade enVisionmath students started out at a lower math performance than 4th grade control students, they later surpassed control students at post-testing."[64] No significant differences from controls were found at grade two. The WWC combined grades two and four without presenting the differences separately.

Student Outcomes Favor Treatment or Are Not Fully Reported

The usefulness of effect-size estimates from RCTs of curricular interventions are no better than the fairness and accuracy of the student outcome measures from which the effect sizes are derived. The NRC report identifies curricular validity as one of the criteria for selecting measures of student outcomes.[65]

Student Outcome Assessments Aligned to Favor the Treatment. The *WWC Procedures and Standards Handbook* requires that these assessments meet traditional criteria of face validity and reliability. The criteria also address overalignment:

> When outcome measures are closely aligned with or tailored to the intervention, the study findings may not be an accurate indication of the effect of the intervention. For example, an outcome measure based on an assessment that relied on materials used in the intervention condition but not in the comparison condition (e.g., specific reading passages) likely would be judged to be over-aligned.[66]

However, the WWC submission form describing outcome measures makes no direct mention of overalignment of outcome assessments that favor the treatment group.[67]

We think the WWC does a good job of requesting and reviewing information to evaluate assessments in terms of face validity and reliability. Therefore, our focus is on identifying studies with the potential for overly aligned assessments, skewed to favor the math content of the treatment intervention.

In the absence of item analyses of each assessment, we use two proxies for an assessment favoring the treatment intervention: (1) there is a presumptive bias because the assessment was created by the developer; or (2) a general standardized assessment was explicitly overaligned by selecting a subset of items conforming to the treatment curriculum.

Table 9 describes the five studies of mathematics identified with likely assessments favoring the treatment intervention. Four of the assessments that appear

Table 9. RCT Mathematics Studies for Which an Assessment Likely Favors the Treatment Curriculum

Intervention	Study	Test Favoring Treatment	Effect-Size Advantage Compared with Other Tests of the Same Students
Accelerated Math (Elementary)	Ysseldyke and Bolt	STAR Math test developed by Renaissance Learning, the developer of Accelerated Math	STAR Math effect size 0.06 greater than with Terra Nova
enVisionMATH	Resendez and Azin	Group Mathematics Assessment and Diagnostic Evaluation (GMADE): Concepts and Communication subtest developed by Pearson, the developer of enVisionMath	GMADE effect size 0.09 greater than average of other three tests
Investigations in Number, Data, and Space	Gatti and Giordano	Group Mathematics Assessment and Diagnostic Evaluation (GMADE): Concepts and Communication subtest developed by Pearson, the developer of Investigations	Publisher developed GMADE is only test
Peer-Assisted Learning Strategies	Fuchs et al.	Stanford Achievement Test (SAT): Subset of aligned items with Peer-Assisted Learning Strategies	SAT aligned test effect size 0.17 greater than with SAT: unaligned items
Accelerated Math (Middle)	Ysseldyke and Bolt	STAR Math test develped by Renaissance Learning, the developer of Accelerated Math	STAR Math effect size 0.26 greater than with Terra Nova

Source: WWC intervention reports.

to favor the treatment curriculum were created by that curriculum's developer. The fifth had an assessment explicitly realigned to conform better to the treatment content. Additionally, four of the potentially overly aligned assessments were paired with other assessments in the same study, and in each case, the assessment likely favoring the treatment yields a larger positive treatment effect size (final column, Table 9).

Fuchs et al. illustrate how outcomes are sensitive to alignment with treatment curriculum content (Table 10).[68] Unlike the other assessments mentioned, Fuchs et al. explicitly identified developing an assessment to align with the Peer-Assisted Learning Strategies (PALS) curriculum. The assessment used was a modification of the Stanford Achievement Test (SAT), created by asking teachers who used the PALS curriculum to identify items that aligned with that curriculum.

The teachers identified 72 aligned and 22 unaligned items on the SAT (Table 11). The aligned SAT items yielded a statistically significant 0.14 effect size for PALS. By comparison, the unaligned SAT items yielded a statistically insignificant –0.03 effect size for PALS. These results suggest that a researcher can improve estimated RCT treatment effect sizes by administering assessments that favor the content taught by the treatment intervention.

Finally, the WWC reported the average of the four RCT studies' effect sizes from the independent and publisher-influenced assessments as their bottom-line measure of the treatment's effectiveness. The problem with this is that the WWC reported its effectiveness measure as an average of the effect sizes of both types of assessments in these four studies, instead of presenting just the effect sizes of the neutral assessment.

WWC Failure to Report Significant Interactions with Student Characteristics. Math curricula vary in content and approach in ways that might differentially

Table 10. Comparison of Student Outcomes on Aligned and Unaligned SAT Assessments with the PALS Intervention

Stanford Achiement Test	Mean difference	Effect Size	p-value
Aligned items with PALS intervention	2.49	0.14	<0.02
Unaligned items with PALS intervention	–0.1	–.03	0.75

Source: WWC Intervention Report: Peer-Assisted Learning Strategies.

Table 11. RCT Mathematics Studies with Identified Student Characteristics Interacting with Outcomes

Study	Author	Examples of Student Characteristics Showing Significantly Greater Treatment Gains Relative to Similar Control Groups
Odyssey Math	DiLeo	SES
enVision Math	Resendez and Azin	Racial/ethnic minorities, females, high-ability students on pretest
Saxon Math and Scott Foresman–Addison Wesley	Agodini, Harris, Thomas, Murphy, and Gallagher	School at the middle or highest third on fall achievement; schools with higher than average percentage of school-lunch eligibility
Saxon Math and Investigations	Agodini, Harris, Thomas, Murphy, and Gallagher	Schools with higher than average percentage of school-lunch eligibility
Investigations	Gatti and Giordano	The fifth-grade Investigations: African American, Caucasian, both free and reduced-priced lunch, female, English proficient, and both higher- and lower-achieving students
Peer-Assisted Learning Strategies	Fuchs, Fuchs, Yazdian, and Powell	Students with disabilities
Scott Foresman–Addison Wesley	Resendez and Azin	Free and reduced-price lunch

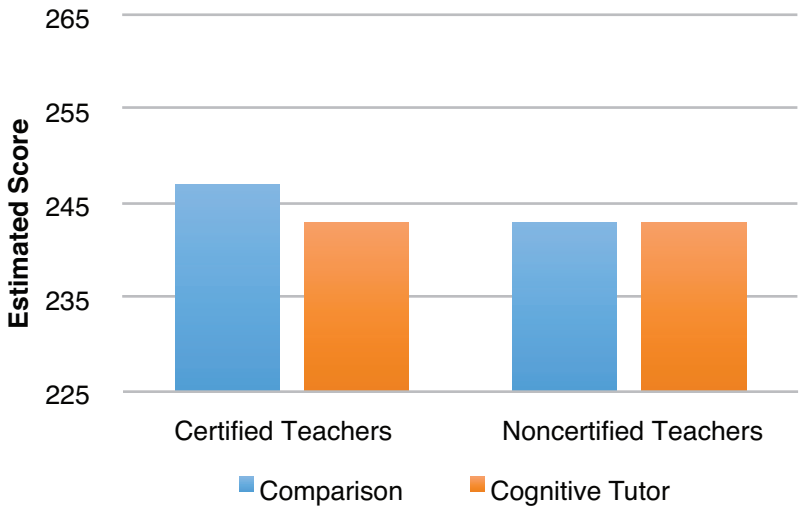
Note: One study of DreamBox examined student interactions with net treatment-comparison gains and found no effects. Significance was based on study authors' reported findings and not on IES review that includes statistical adjustments for clustering.
Source: DiLeo, "A Study of a Specific Language Arts and Mathematics Software Program"; Resendez and Azin, *A Study on the Effects of Pearson's 2009 enVisionmath Program;* Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula*; Gatti and Giordano, *Investigations in Number, Data, and Space Efficacy Study*; L. S. Fuchs et al., "Enhancing First-Grade Children's Mathematical Development with Peer-Assisted Learning Strategies"; and Resendez and Azin, *2005 Scott Foresman–Addison Wesley Elementary Math Randomized Control Trial.*

affect students at different performance levels or backgrounds. Accordingly, the NRC criteria for evaluating curricula effectiveness recommend: "Disaggregate data by gender, race/ethnicity, socioeconomic status (SES), and performance levels, and express constraints as to the generalizability of the study."[69]

Unlike the NRC recommendation, the WWC standards do not directly encourage breaking out student test results by student characteristics. The WWC standards state procedures only for aggregation of data across student subgroups.[70]

Because the WWC intervention reports do not present findings about interactions of outcomes with student characteristics, we reviewed the available study reports. Among the 18 studies for which we could examine the original reports, the authors of 7 studies identified student subgroups with statistically significant interaction effects in relation to the corresponding comparison treatment (Table 11). These interactions were found across a range of student characteristics, including race and ethnicity, those receiving free and reduced-price lunch, English-proficient students, and

Figure 2. RCT Study of the Effects on Student Outcomes Between Cognitive Tutor and Control Interventions, by Certified and Uncertified Teachers



Source: Cabalo and Vu, *Comparative Effectiveness of Carnegie Learning's Cognitive Tutor Algebra I Curriculum.*

students with disabilities. Having access to the data on effects for these population subgroups is important for school systems trying to understand a curriculum's effectiveness for their own students.

WWC Failure to Report Significant Interactions with Teacher/Teaching Characteristics. The NRC review stressed the importance of examining teacher factors in evaluations of curricula effectiveness:

> Many evaluation studies devoted inadequate attention to the variable of teacher quality. . . . Hardly any of the studies actually do anything analytical, and because these are such important potential confounding variables, this presents a serious challenge to the efficacy of these studies.[71]

The *WWC Procedures and Standards Handbook* discusses teacher effects under a larger category of confounding factors, such as the teachers in the intervention group having master's degrees and the comparison group having only bachelor's degrees.[72]

In practice, the WWC mathematics intervention reports did not discuss interactions of outcomes with teacher characteristics. However, we identified four RCT math studies from the original study reports that estimated interactions between characteristics of teachers with the intervention and control curricula.

For example, Agodini et al., which includes three of the four studies contained in the WWC, found that differences in the comparative effectiveness of curriculum interventions pairs (Table 7) were associated with differences in teachers' respective use of student-centered or teacher-directed instructional approaches.[73] This is consistent with the overall effectiveness results, where the two relatively teacher-directed curricula outperformed the two student-centered curricula.

A second illustration of teacher characteristics interacting with the intervention is Cabalo and Vu, which identified significant effects of teacher certification on the differences in student outcomes between the Cognitive Tutor and the comparison group (Figure 2).[74] With respect to certified teachers, the controls outperformed the Cognitive Tutor teachers. For noncertified teachers, the controls and Cognitive Tutor outcomes were equivalent.

The authors theorize that these differences occured because certified teachers are familiar with the

comparison curriculum, giving them an advantage over the Cognitive Tutor teachers, who are teaching this curriculum for the first time. However, noncertified teachers may not be familiar with either curriculum. Thus, the Cognitive Tutor teacher interaction findings about certification may well be consistent with the threat of conducting RCTs during the first year of the treatment intervention's implementation. The WWC did not report these findings, even though this information would be useful to school systems deciding on a curriculum's effectiveness and its emphasis on particular mathematics teaching approaches.

Outdated Curricula

A useful study should be current. That is, if WWC customers purchased the current curriculum, it would be the same or similar in content, structure, and instructional approach to the edition that the RCT study evaluated.

This condition is not met when the current edition of a curriculum has changed in a significant way, including major alterations in topic structure, content, pedagogy, assessments, or use of technology or technology content. In such cases, the effectiveness results of even the most sophisticated and accurate RCT study may be outdated and unhelpful for a customer in gauging the current curriculum's potential effectiveness.

The NRC report did not address the issue of outdated curricula. This is not surprising, given that their "charge was to evaluate the quality of the evaluations of the 13 mathematics curriculum materials supported by the National Science Foundation (NSF) . . . and 6 of the commercially generated mathematics curriculum materials."[75] These curricula were developed largely during the 1990s, and hence, the NRC review was of comparatively current curricula.

This current review uses any one or more of three criteria to establish an out-of-date curriculum. First, it was published in 2000 or before. The WWC intervention report typically contains the date of the RCT and information about the edition of the curriculum, but our measure also requires a date for the edition, which is often not given. When the date of the edition is available, that date is used. Otherwise, the date shown for the RCT study in the field is used, or if that is not available, the date of the published study is used.

Second, the version is no longer published, which is verified by checking with the publisher for availability of each curriculum. Third, a newer version based on career- and college-ready state standards, such as the Common Core, is available, which is verified by visiting the publisher's website of curricula's publisher.[76]

The overall pattern of the data on study currency is as follows:

- Five studies used curricula from before 2001, with the earliest going back to 1981. We presume these curricula are out of date with current state standards and the application of educational technology.

- One study published after 2000 refers to a curriculum that the publisher has replaced and no longer currently supports.

- Among the 21 remaining studies, 13 have published an edition aligning with the Common Core or other new college- and career-ready state standards. Because all the WWC math studies were published before the development of the Common Core standards, we would expect that an edition aligned with the Common Core standards would not only differ in topic content by grade, but also in the emphasis on the eight Common Core curriculum processes that stress deeper mathematical thinking.[77] Note that some of these non–Common Core curricula may be used in the few states that have not adopted the Common Core or its equivalent at the state standards.

- In total, this means that 8 of the 27 studies with the intervention curricula were published after 2000, are currently available, and have no available Common Core edition.

Summary

This analysis examined the usefulness of 27 studies of mathematics curricula. The curricula were embedded within 23 RCTs, which the WWC concluded were acceptable evidence. We reviewed 12 threats other than selection bias for each of the 27 RCT studies. Eight of these threats arise from RCT design and implementation problems, and four arise from the WWC's reporting of the RCTs. Tables A3 and A4 summarize the presence of threats we documented for each intervention and include the following:

- In 12 of the 27 RCT studies (44 percent), the authors appear to have an association with the developer. Four of these 12 also have comparable independent studies, which all have a smaller effect size than the study associated with the developer.

- In 23 of 27 studies (85 percent), implementation fidelity is threatened, because the RCT occurred in the first year of the curriculum implementation, so the teacher was likely teaching the curriculum for the first time. The NRC warns that it may take up to three years to implement a substantially different curricular change.

- Eight of the 14 RCT curriculum studies measuring implementation indicate low implementation fidelity. Six studies also found a relationship between the degree of implementation fidelity (for example, number of curricular objectives completed) in the intervention classrooms and student outcomes (Table A1).

- In 15 of 27 studies (56 percent), the comparison curricula are either never identified or, if identified, outcomes are reported for a combined two or more comparison curricula. Without understanding the characteristics of the comparison, we cannot interpret the intervention's effectiveness.

- In eight of nine studies for which the total time of the intervention is available, the treatment time differs substantially from that of the comparison group (Table A2). In six of the eight studies, the intervention time is longer.

- In 19 of 20 studies, a curriculum covering two or more grades does not have a longitudinal cohort to measure cumulative effects across grades. This makes it impossible to look at the cumulative effects of a curriculum intervention.

- Thirteen of 27 studies (48 percent) do not cover a sufficiently broad range of grades to provide a clear picture of curriculum effectiveness. These include 7 of 12 elementary studies, and 6 of 10 middle school studies.

- In 5 of 27 studies (19 percent), the assessment was designed by the curricula developer and likely is aligned in favor of the treatment. Four studies administered a neutral assessment not favoring the treatment, and in all cases, the assessment favoring the treatment yielded larger effect sizes for the treatment compared with the more neutral assessments. The problem is that the WWC reported its effectiveness measure as an average of both assessments' effect sizes in these four studies, instead of presenting just the the neutral assessment's effect size (Table 9).

- In 19 of the 27 studies (70 percent), the RCTs are carried out on outdated curricula, meaning it was developed before 2000, is no longer currently supported, or has been revised to align with new college- and career-ready standards, such as Common Core.

In addition, several studies contained evidence about interactions of outcomes with student characteristics, teacher instructional approaches, or teacher characteristics, but the WWC does not report these interactions. These findings might have provided important information about the applicability of treatments to students or teachers with particular characteristics.

Moreover, the magnitude of the error generated by even a single threat is frequently greater than the average

Table 12. Frequency of Threats, by WWC Math Intervention RCT Studies

Number of Threats	Elementary Frequency	Middle/Secondary Frequency
1	–	1
2	–	4
3	–	1
4	–	3
5	3	2
6	4	3
7	3	–
8	1	–
9	1	1
10	–	–
Total	**12**	**15**

Source: Original studies when available; and Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0*. See Tables A3 and A4 for details.

effect size of an RCT treatment. For the seven studies that had statistically significant and positive treatment effect sizes, the median effect size was 0.15. By comparison, the following three sources of threat errors were quantifiable in terms of effect size, and each had magnitudes of at least 0.15:

- The interventions where a study was associated with the curriculum publisher had a 0.21 average advantage in effect size.

- Saxon Math had a 0.22 range in effect sizes, depending on the choice of one of three comparison curricula.

- When the assessment favored the intervention, the intervention curriculum had a 0.15 average advantage in effect size.

Overall, the data in the Appendix show that all the RCTs in the sample have serious threats to their usefulness. This raises the issue of whether any of the RCTs provide sufficiently useful information for consumers wishing to make informed judgments about which mathematics curriculum to purchase.

Table 12 displays the frequency distribution of threats across all the elementary and secondary studies. Every elementary grade study has at least four identified threats. One middle/secondary school study has only one threat. All the rest have at least two threats. One reason for the lower number of threats at the high school level is that the studies were designed primarily for a single year of an algebra or geometry curriculum and, hence, are not subject to grade coverage weaknesses.

Even those studies with only one or two threats hold considerable uncertainty in their usefulness for curricular choices. Table 13 displays the five studies identified at the middle or high school level with one or two threats and nonsignificant effect sizes. In two studies, the curriculum was in the first year of implementation. In two other studies, the control group is ambiguous. Moreover, the first two RCTs in Table 13 were conducted before the mid-1990s, and the curricula are not current.

One Cognitive Tutor study is developer associated, and this study displayed the only positive effect size out of the four Cognitive Tutor studies. Finally, one Cognitive Tutor study has only one threat (the lack of a specific control group), but this creates ambiguity in interpreting effect-size estimates from the RCT.

Table 13. WWC Math Studies with One or Two Threats

Study	Author	Effect Size	Threats
Transition Math	Baker	–0.35 (nonsignificant)	• First year of implementation • Not current (1994)
University of Chicago Mathematics Project (Algebra)	Peters	–0.14 (nonsignificant)	• First year of implementation • Not current (1991)
Cognitive Tutor Algebra	Ritter, Kulikowich, Lei, McGuire, and Morgan	0.38 (nonsignificant)	• Developer associated • First year of implementation
Cognitive Tutor Algebra	Campuzano, Dynarski, Agodini, and Rall	–0.16 (nonsignificant)	• No specific named comparison group
Cognitive Tutor	Pane, McCaffrey, Slaughter, Steele, and Ikemoto	–0.19 (nonsignificant)	• Poor implementation and fidelity of implementation associated with improved outcomes • No specific named comparison group

Note: Based on joint results of Cognitive Computer and Larson Algebra.
Source: Original studies when available; and Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0*.

We conclude that none of the RCTs provides sufficiently useful information for consumers wishing to make informed judgments about which mathematics curriculum to purchase.

Part III:
Observations and Recommendations

This report examines the 27 mathematics RCT curriculum studies contained in the WWC in December 2014. We focused on 12 threats to the usefulness of the study information to educators: eight of the threats are because of the design and implementation of the RCTs, while the other four stem from reporting decisions made by the WWC. We conclude that each of the 27 curriculum studies contains enough significant threats to seriously jeopardize the usefulness of the information reported by the WWC.

Our conclusion suggests three questions. We explore possible answers to these questions in turn before offering five policy recommendations to the WWC, the IES and the OMB.

Is it possible to carry out a robust, timely, economical, and generally threat-free RCT that provides educators with useful information about the effectiveness of a mathematics curriculum? One cautionary consideration is that a teacher-implemented mathematics curriculum is a complex intervention, and schools and classrooms are potentially unstable environments that evaluators cannot control. Randomized control studies that attempt to examine the effectiveness of mathematics curricula are therefore particularly vulnerable to threats related to the intervention's design and implementation. Because of this complexity, the challenge of overcoming the potential threats defined in this report could require a lengthy, difficult, and expensive process, which is fraught with potential compromises.

Hypothetically, such a robust RCT study of a math curriculum might be selected, designed, implemented, analyzed, and published by researchers and then reviewed and reported by the WWC. However, there is substantial time between the release of a curriculum and the release of a useful WWC evaluation report. If a new version of the curriculum supplants the original by the WWC report release, educators would find little use for the report.

Assuming that the evaluators worked quickly, a relatively simple example would be an RCT designed to evaluate a curriculum used in a one-year course, such as Algebra I. The use of a single-year course eliminates the threat of having to sample two or three years of a multiyear curriculum to adequately assess its effectiveness, which is time intensive and expensive.

The algebra study would have to address other threats, such as ensuring that the teachers who use the intervention have taught it for at least one to two years before the study year. This requires overcoming multiple implementation challenges. For example, the process of randomization has to be carried out a year before the data are collected. This is expensive and politically and technically difficult, especially in a high school where students are often unsure about the courses they are going to take and parents are concerned about their children being used in experiments.

This is a discouraging picture, and we have considered only one of the eight design and implementation potential flaws. Yet addressing each of the potential flaws is possible if enough resources are available, and for each of the threats we examined, we can find at least one RCT that addressed it. Addressing all the threats together creates a more challenging problem.

Is this report's conclusion applicable to other curriculum content areas and to other forms of education intervention? Our answer to the first part of this question is almost certainly yes. We would be very surprised if the RCTs of similar curriculum studies in other areas, such as literacy and science, did not contain the same

threats as mathematics curricula RCTs. The practices of the WWC are surely the same, as are all the conditions and issues attendant to evaluating complex curricular interventions using RCTs.

The second part of the question has to do with interventions other than curricula. They include early childhood, some reading interventions, higher education, K–12 dropout behavior, and other sectors such as the training programs in the Labor Department. Just as with the curriculum studies, the implementation of most of these interventions is complex, and the environment challenging. In this regard, the conditions are similar to the curriculum studies, and similar threats may occur.

What steps should be taken to provide information and support to educators making decisions about curricula and other educational interventions? We do not know for certain, but we have some observations and thoughts that might prompt a conversation. First, the effect sizes from the 27 curriculum studies are quite small; almost all are under 0.2, and few are statistically significant. This could be because of the influence of the various threats, but it also might be because of the conservative nature of the US education system.

Only a few large publishers dominate the American market, and the US standards-based system asks states and local districts to align their content and assessments with the state standards. These factors do not reward diverse curricula and experimental approaches to instruction, and they tend to homogenize the published curricula. We should expect to find only small effects in these studies, and therefore threat elimination is particularly important because nonrandom variation due to threats may be relatively more severe than with large intervention effects. But, ironically, if the effects are tiny and predictable because the theories and practices are similar, there is little reason to carry out the study.

Second, we know from other research that the effectiveness of an intervention depends not only on its quality or the character of the control group but also on its context, including the preparation of the teachers, the social capital of the school, and the fit of the students to the content. These factors can vary substantially.[78] Again, if we anticipate small effects, we might expect the substantial variation in context to have a relatively strong influence on our results.

Third, studies show that successfully reformed districts have often had the same basic curriculum in place for 5 to 10 years, which gives teachers the opportunity to improve their use of the curriculum and alter it to their students' needs.[79] This suggests that it is not the curriculum itself that is effective or ineffective, but rather the teacher's opportunity and use of the curriculum to improve their effectiveness.

Taken together, these factors suggest that we need a more nuanced and thoughtful approach to evaluating curricula. Perhaps we should only compare curricula in settings where we expect to find large effects because of clear differences in pedagogical approach or content emphasis.

Or perhaps curricula should be deliberately designed and constructed to be easily adaptable to various contexts by teachers and others in a school system. If this is the case, a single RCT would not add much useful information. Multiple studies that sample different contexts might be much more valuable.

We might also be more nuanced in our threat analysis, because some threats are more severe than others. The most troublesome threats could be labeled "necessary to overcome" (a red flag), and the others could be viewed as not critical, but still worrisome (a yellow flag).

Something akin to a WWC practice guide might be put together for a particular curriculum, including teachers' experiences, the theory behind the curriculum, and data from RCTs and other studies done in different contexts. The guide could be regularly updated. Alternatively, the WWC could use and summarize the best of the many websites that already collect evaluative data about curricular materials.

In this discussion, we have implicitly assumed that the world of curriculum and teaching has not recently changed and will not change over the next few years. This assumption is wrong on both counts. The WWC was designed 15 years ago, only a few years after the Internet became ubiquitous. Since then, the development and delivery mechanisms of education interventions have changed dramatically.

For example, the delivery of curricula and other

education interventions is steadily migrating from paper platforms to Internet platforms, with multiple and rapidly improving ways of representing and teaching content and skills. This change has significantly affected the development of curricula, as technology enables a wider variety of instructional practices and a move toward rapid prototyping and toward using improvement science to regularly upgrade interventions.[80]

This combination of factors suggests that curricula and other education interventions may no longer be stable for five to seven years. In effect, a particular curriculum may be considered out of date as soon as it is published, which means it cannot be used to develop estimates of its effect sizes compared to other curricula.

These changes are not transient problems, for they already have momentum and will be commonplace in the near future. Rather, they should be seen as new opportunities for experimentation in continuously improving instruction and the innovative use of curricular materials. For example, these general trends have fostered curricular packages that contain evidence-designed materials for teachers to use as part of their year's course of study, and the teachers' work can be supported by Internet platforms that are designed for such "plug and play" materials.

All this opens opportunities for new and inexpensive ways for evaluators to gather data about curricula and curricular materials by using greatly improved data systems and surveying ongoing teacher samples that provide evaluative information about their schools' curricula. We do not pretend to know how all these new factors will play out over the long term, but we are excited about the prospects of new, imaginative, and powerful ways of developing thoughtful judgments on the quality of education interventions. We have five recommendations on how to best seize this opportunity.

Recommendation 1. The IES should review our analyses of the 27 RCT mathematics curriculum reports. It should remove the studies and RCT reports that, in their view, do not provide useful information for teachers and other educators that turn to the WWC for help in deciding their curricula. The IES should make their judgments and their rationale public.

Recommendation 2. The IES should examine the WWC reports of curricula intervention studies outside mathematics that draw on RCTs. It should remove the studies and RCTs that, in its view, do not meet the standard of providing useful information for educators. It should make public its judgments and the rationale behind the judgments.

Recommendation 3. The IES should review a representative sample (at least 25 percent) of all the noncurricula education intervention studies that use RCTs in the WWC. The review should include the same criteria and standards as in recommendations 1 and 2. Studies that do not meet these standards should be removed from the WWC.

If a significant percentage of studies in the sample do not meet the standards (for example, more than 25 percent), the IES should review all RCTs and associated intervention studies currently posted on the WWC. All studies and RCTs that do not meet their standards should be removed from the WWC. The results of the review should be made transparent and public.

Recommendation 4. This recommendation has two parts. First, the IES should immediately create an internal expert panel of evaluators, curriculum experts, and users (such as teachers and administrators) to consider how to improve the current WWC criteria in the short term and develop standards for reviewing RCTs, curriculum, and noncurricula intervention studies that include education RCTs. The panel should prepare standards to be used in recommendations one through three. This commission's activities and conclusions should be public and transparent.

Second, taking into consideration the significant changes in the education world and in the availability of useful data, the IES and OMB should support an ongoing, five-year commission of experts convened by the NRC or the National Academy of Education. The commission would consider effective and useful evaluation and improvement systems for educational

materials and practices for the future. They should also consider how this system might be developed and supported and what the appropriate role of the federal government should be in designing, creating, and administering this system. The panel report should be made publicly available and free.

Recommendation 5. Finally, we recommend that the OMB support a three-year study by a commission of unbiased experts and users convened by the NRC to look at the usefulness of RCT studies in parts of the government outside of education. We see no reason to expect that RCTs funded out of the Labor Department, HUD, Human Services, Transportation, USAID, or parts of the Education Department other than the WWC (for example, I3) would be immune to the flaws we find in the RCTs in the WWC. The activities and conclusions of this panel should be transparent and public.[81]

About the Authors

Alan Ginsburg and Marshall S. Smith are retired federal officials from the US Department of Education.

Acknowledgments

We thank the Spencer Foundation and the Carnegie Foundation for the Advancement of Teaching for their support and George Bohrnstedt, Robert Boruch, Adam Gamoran, Robert Hauser, Kelsey Hamilton, Fredrick Hess, Eugenia Kemble, and Nat Malkus for their thoughtful comments on earlier drafts. We also thank the American Enterprise Institute for its support in publishing our paper.

Appendix

The appendix is available online.

Notes

1. What Works Clearinghouse, "About the WWC," Institute of Education Sciences, http://ies.ed.gov/ncee/wwc/aboutus.aspx.

2. Bob Boruch and Erling Boe point out that "interest in our topic of use of dependable evidence has early origins, notably in John Graunt's declarations in the 17th Century." Graunt's *Nature and Political Observations Made upon the Bills of Mortality* addressed "to what purpose tends all this laborious bustling and groping? To know the number of . . . people, fighting men, teeming women, what years are fruitful." Graunt's response includes to improve government "by the knowledge whereof trade and government may be made more certain and regular . . . so trade might not be hoped for where it is impossible." In the modern era, the WWC aims to make accessible online useful evidence to customers at all governmental levels. R. Boruch and E. Boe, "On 'Good, Certain, and Easy Government:' The Policy Use of Statistical Data and Reports," in *Effective Dissemination of Clinical and Health Information*, ed. L. Sechrest, T. Backer, E. Rogers, T. Campell, and M. Grady (Public Health Service, US Department of Health and Human Services, 1994); and John Graunt, *Nature and Political Observations Made upon the Bills of Mortality* (1662), 96.

3. R. A. Fisher and J. Wishart, "The Arrangement of Field Experiments and the Statistical Reduction of the Results," *Imperial Bureau of Soil Science Technical Communication* 10 (1930).

4. The potential usefulness of randomized experiments in education was highlighted in the mid-1980s by the Tennessee Class Size Experiment. This was a particularly well-run experiment with a relatively simple intervention (lower class size) that produced positive results. The study was reviewed later by Frederick Mosteller from Harvard, one of the world's best-known statisticians, and given a "thumbs up." See Frederick Mosteller, "The Tennessee Study of Class Size in the Early School Grades," *The Future of Children: Critical Issues for Children and Youths* 5, no. 2 (Summer/Fall 1995), https://www.princeton.edu/futureofchildren/publications/docs/05_02_08.pdf.

5. There is an exception to this sentence. If the intervention and control samples are both drawn randomly from a very well-defined population, then the RCT might produce evidence that is externally valid for the well-defined population. None of the RCT studies in the WWC of mathematics curriculum had random samples drawn from a well-defined population. Of course, the RCT would have to also be internally valid.

6. William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin, 2002).

7. A. Jaciw and D. Newman, "External Validity in the Context of RCTs: Lessons from the Causal Explanatory Tradition," paper presented before the Society for Research on Educational Effectiveness, 2011.

8. D. Campbell, "Reforms as Experiments," *American Psychologist* 24 (1969): 409–29.

9. The NRC report states: "However, curricular effectiveness cannot be established by a single scientifically valid study; instead a body of studies is needed." National Research Council, *On Evaluating Curricular Effectiveness: Judging the Quality of K–12 Mathematics Evaluations*, ed. Jere Confrey and Vicki Stohl (Washington, DC: National Academies Press, 2004), 191, http://www.nap.edu/catalog/11025.html.

10. Nancy Potischman, a nutritional epidemiologist at the National Cancer Institute, provides an example of how health research stresses replication: "We never take any one study to be the answer to anything. . . . Only if the same results come up in multiple studies across multiple populations . . . then you might think that, yes, this food might be important." S. Levingston, "Looking for That Fruit or Vegetable That Might Prevent Cancer?" *Washington Post,* February 16, 2015.

11. Campbell, "Reforms as Experiments."

12. P. Craig et al., "Developing and Evaluating Complex Interventions: The New Medical Research Council Guidance," *BMJ* 337 (2008).

13. We did not examine studies of the mathematics curricula other than the 27 RCTs. There may be powerful and justifiably convincing evidence from the non-RCT studies about the effectiveness of the math curricula evaluated on the WWC website.

14. See Institute of Education Sciences, "What Works Clearinghouse," http://ies.ed.gov/ncee/wwc/.

15. National Research Council, *On Evaluating Curricular Effectiveness.*

16. Jere Confrey, "Comparing and Contrasting the National Research Council Report 'On Evaluating Curricular Effectiveness'

with the What Works Clearinghouse Approach," *Educational Evaluation and Policy Analysis* 28, no. 3 (2006): 195–213.

17. A. H. Schoenfeld, "What Doesn't Work: The Challenge and Failure of the What Works Clearinghouse to Conduct Meaningful Reviews of Studies of Mathematics Curricula," *Educational Researcher* 35, no. 2 (2006).

18. A. Cheung and R. Slavin, "The Effectiveness of Educational Technology Applications for Enhancing Mathematics Achievement in K–12 Classrooms: A Meta-Analysis," *Educational Research Review* 9 (2013): 88–113.

19. In total, the WWC reported on 27 RCT math studies in December 2014. By comparison, when the NRC published its study on curricula effectiveness in 2004, it could not find a single RCT among the 63 studies of the mathematics curricula meeting their evaluation standards. National Research Council, *On Evaluating Curricular Effectiveness,* 63.

20. For example, Baker compared the Expert Mathematician intervention with the Transition Mathematics intervention, and the comparison is listed twice, once under the effectiveness of the Expert Mathematician and again under the effectiveness of the Transition Mathematics. These are two separate intervention reports, and each comparison is listed as a separate study. J. J. Baker, "Effects of a Generative Instructional Design Strategy on Learning Mathematics and on Attitudes Towards Achievement," *Dissertation Abstracts International* 58, no. 7 (1997).

21. During our review, we did identify some instances of surprisingly large attrition in studies that still met WWC review without reservations. For example, Cabalo et al. had an attrition rate of 33 percent from the randomized assignment but was still rated as fully meeting WWC criteria without reservations. Cabalo, and M. T. Vu, *Comparative Effectiveness of Carnegie Learning's Cognitive Tutor Algebra I Curriculum: A Report Of A Randomized Experiment in the Maui School District*, Empirical Education Inc., 2007.

22. The What Works Clearinghouse Glossary defines the WWC acceptance standards as follows: Meets WWC group design standards without reservations: "The highest possible rating for a group design study reviewed by the WWC. Studies receiving this rating provide the highest degree of confidence that an observed effect was caused by the intervention. Only well-implemented randomized controlled trials that do not have problems with attrition may receive this highest rating." Meets WWC group design standards with reservations: "The middle possible rating for a group design study reviewed by the WWC. Studies receiving this rating provide a lower degree of confidence that an observed effect was caused by the intervention. Randomized controlled trials that are not as well implemented or have problems with attrition, along with strong quasi-experimental designs, may receive this rating." What Works Clearinghouse, "Glossary," http://ies.ed.gov/ncee/wwc/glossary.aspx.

23. We did not recover nine of the original full studies supporting the 27 RCTs. The nonrecovered studies typically were dissertations or were contracted by a curriculum publisher that did not display the full original study report on its website.

24. In using the word "criteria," we adopted the language for looking at potential threats used in the NRC report, *On Evaluating Curricular Effectiveness.* The WWC instead uses the term "standard," which may connote a firmer threshold of acceptable threat levels for a study to be accepted.

25. A strong theory provides reasons for choosing an intervention to study, expects positive results, and often provides a road map for understanding how well the intervention is being implemented. This in turn provides insight about the results if they are not positive. The NRC report argues that a replication of the RCT should be carried out before the RCT is seen as providing strong evidence. The replication would be conducted in conditions that are as close as possible to the original RCT. If both studies show statistically significant results in the same direction, they provide a much firmer base of statistical support for the evidential claims. In many trials carried out in less complex settings, the person administering the intervention does not know which group uses the intervention and which is in the counterfactual treatment in a single blind experiment. This becomes a double blind when the group itself does not know whether it is in the intervention treatment or the counterfactual treatment. When the researcher knows which group is the intervention group and the group knows they are an "experimental" group, they may behave differently than they would if they did not know.

26. G. J. Whitehurst, *The Institute of Education Sciences: New Wine, New Bottles*, American Educational Research Association, 2003 Annual Meeting Presidential Invited Session, April 22, 2003, http://ies.ed.gov/director/pdf/2003_04_22.pdf.

27. National Research Council, *On Evaluating Curricular Effectiveness*; and Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0,* http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf.

28. Among the four threats that arise from WWC reporting of the findings, three threats occur when the WWC does not report details about outcomes available from the study, including not reporting outcomes by grade, student interactions, or teacher interactions. The fourth such threat occurs when the WWC does not report that a curriculum is likely out of date.

29. National Research Council, *On Evaluating Curricular Effectiveness,* 61.

30. Note that this cannot happen if there is randomization of both students and teachers.

31. A. Franco, N. Malhotra, and G. Simonovits, "Publication Bias in the Social Sciences: Unlocking the File Drawer," *Science* 345 no. 6203 (September 2014): 1502–1505.

32. J. Valentine et al., "Replication in Prevention Science," *Prevention Science* 12 (2011): 103–17.

33. Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook;* and Institute of Education Sciences, *What Works Clearinghouse Reporting Guide for Study Authors,* http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=235.

34. In a footnote, the authors of an Accelerated Math study included the following acknowledgement of developer Renaissance Learning's involvement in the paper: "The authors acknowledge the assistance of Michael Patterson of the Research Department at Renaissance Learning with data collection and manuscript editing." Gatti and Giordano note that "Pearson contracted Gatti Evaluation, an independent research consulting company, to conduct a longitudinal evaluation of the impact of the *Investigations in Number, Data, and Space ©2008* (*Investigations*) mathematics curriculum on student mathematics achievement and attitudes." Jim Ysseldyke and Daniel M. Bolt, "Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement," *School Psychology Review* 36, no. 3 (2007): 453–67, http://questgarden.com/76/55/0/090212083730/files/progress%20monitoring%20math.pdf; and G. Gatti and K. Giordano, *Investigations in Number, Data, & Space Efficacy Study: Final Report*, Gatti Evaluation Inc., 2010, http://www.gattieval.com/IND&S_Study_Report.pdf.

35. An evaluation company where association between the evaluator and publisher was not explicit but considered quite likely was PRES Associates' four RCT studies. PRES Associates has on its website the following nondisclosure statement to justify withholding clients' names: "Examples of recent projects conducted by the staff of PRES Associates include: . . . Several large-scale randomized control trials evaluating the effectiveness of reading and math curriculum materials. These studies have been conducted for private organizations and publishers (client names withheld for confidentiality)." PRES Associates Inc., "Expertise," http://presassociates.com/expertise/. We attempted to contact an author to obtain more information about the nature of the study's relationship to the curricula developers, but we received no response from PRES Associates. On November 17, 2014, we emailed Miriam Resendez, the vice president of PRES Associates, asking about her and PRES Associates' relationship with the curricula developers evaluated under PRES Associates. No response from Ms. Resendez was received. We treated the four studies conducted as developer associated. This experience suggests that the WWC may want to include evaluator-association information routinely as part of their intervention reports, a practice IES already follows with their own funded evaluation studies.

36. Note that in Table 4, the entries for SFAW in the "Yes, an Association" column and for Cognitive Tutor Algebra in the "No Association" column are averages across two or more studies.

37. National Research Council, *On Evaluating Curricular Effectiveness,* 6.

38. Ibid., 114.

39. Institute of Education Sciences, *What Works Clearinghouse Reporting Guide for Study Authors*.

40. Peter H. Rossi, Howard E. Freeman, and Mark W. Lipsey, *Evaluation: A Systematic Approach* (Thousand Oaks, California: Sage, 1999), 238.

41. National Research Council, *On Evaluating Curricular Effectiveness,* 61.

42. John F. Pane et al., "An Experiment to Evaluate the Efficacy of Cognitive Tutor Geometry," *Journal of Research on Educational Effectiveness* 3, no. 3 (2010): 254–81, http://www.rand.org/pubs/external_publications/EP20100057.html.

43. Judy DiLeo, "A Study of a Specific Language Arts and Mathematics Software Program: Is There a Correlation Between Usage Levels and Achievement?" (doctoral dissertation, Indiana University of Pennsylvania, May 2007), http://dspace.iup.edu/bitstream/handle/2069/40/Judy+DiLeo.pdf?sequence=1.

44. To illustrate, let's presume the students in the grade before the study grade use both Odysseys, so a student might have had

Odyssey Math and Reading in third grade and then Odyssey Math or Reading in fourth grade. The student would then already have been exposed to Odyssey Math, so the comparison is in some ways comparing Odyssey Math for one year with Odyssey Math for two years.

45. Agodini et al. presented results for 110 elementary schools that had been randomly assigned to one of four conditions: Investigations in Number, Data, and Space (28 schools), Math Expressions (27 schools), Saxon Math (26 schools), and Scott Foresman–Addison Wesley Elementary Mathematics (29 schools). Roberto Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula: Findings for First and Second Graders,* National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, October 2010, http://ies.ed.gov/ncee/pubs/20114001/pdf/20114001.pdf.

46. Miriam Resendez and Mariam Azin, *A Study on the Effects of Pearson's 2009 enVisionmath Program,* PRES Associates Inc., September 2008, http://www.pearsoned.com/wp-content/uploads/envisionmath_efficacy_study_year1_final_report.pdf.

47. Ibid., 40.

48. Peggy C. Kirby, *I CAN Learn in Orleans Parish Public Schools: Effects on LEAP 8th Grade Math Achievement, 2003–2004*, Ed-Cet Inc., October 2006, http://www.icanlearnresults.com/pdf/Orleans%208th%20grade.pdf.

49. Kirby also had an association with the I CAN Learn developer, illustrating the importance of informing WWC users of such ties.

50. Another example is that the effects of a program such as Investigations in Number, Data, and Space, which stresses student-centered, real-world problem solving, may depend on whether the program is compared to a similar student-centered program or against a teacher-centered one. In this instance, the study effects also would be affected by the two curriculas' degrees of alignment to the content of the outcome measure. A third example is the simple idea that a curriculum intervention may look very good if the comparison curriculum is very weak and vice versa. The comparison curriculum seems to be as important as the experimental intervention in influencing the effects sizes reported from the study.

51. National Research Council, *On Evaluating Curricular Effectiveness,* 106.

52. Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook,* 5.

53. Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula.*

54. Ibid., 77.

55. National Research Council, *On Evaluating Curricular Effectiveness.*

56. Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook.*

57. L. S. Fuchs et al., "Enhancing First-Grade Children's Mathematical Development with Peer-Assisted Learning Strategies," *School Psychology Review* 31, no. 4 (2002): 569–83.

58. Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula.*

59. National Research Council, *On Evaluating Curricular Effectiveness,* 107.

60. Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook,* 17.

61. A second study, by Miriam Resendez and Mariam Azin, presented two years of longitudinal data, but the WWC only reported the results for the first. Like Gatti and Giordano, the second-year results in several cases reversed the findings from the first year. Resendez and Azin, *A Study on the Effects of Pearson's 2009 enVisionmath Program*; and Guido G. Gatti and Kate Giordano, *Investigations in Number, Data, & Space Efficacy Study: Final Report*, Gatti Evaluation Inc., 2010, http://www.gattieval.com/IND&S_Study_Report.pdf.

62. "Despite high attrition, the difference between the intervention and comparison groups along baseline math achievement was in the range where the study could meet WWC evidence standards with reservations, provided the results were adjusted for the baseline differences." WWC Investigations report.

63. Beth C. Gamse et al., *Reading First Impact Study Final Report,* National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, November 19, 2008, http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20094038.

64. Resendez and Azin, *A Study on the Effects of Pearson's 2009 enVisionmath Program,* 46.

65. National Research Council, *On Evaluating Curricular Effectiveness.*

66. Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook,* 16.

67. Specifically, the WWC submission form states: "For each outcome measure used in the study, describe the measure, how it was collected, how to interpret it, whether it is standardized (if so, using what metric), whether it has been normed (if so, describe the norming population), and, if relevant, who developed it. For non-standardized measures, describe the validity and reliability of the outcome measures based on the study sample." Institute of Education Sciences, *What Works Clearinghouse Reporting Guide for Study Authors.*

68. L. S. Fuchs et al., "Enhancing First-Grade Children's Mathematical Development with Peer-Assisted Learning Strategies."

69. National Research Council, *On Evaluating Curricular Effectiveness,* 7.

70. Ibid., 17.

71. Ibid., 119.

72. "When the characteristics of the units in each group differ systematically in ways that are associated with the outcomes. For example, a small group of teachers in a master's program implements the intervention, whereas students in the comparison group are taught by teachers with bachelor's degrees. If the teacher's education is not a component of the intervention—that is, the intervention does not specify that only master's level teachers can lead the intervention—then it is a potential confounding factor." Institute of Education Sciences, *What Works Clearinghouse: Procedures and Standards Handbook,* 19. "WWC reviewers must decide whether there is sufficient information to determine that the only difference between the two groups that is not controlled for by design or analysis is the presence of the intervention. If not, there may a confounding factor, and the reviewer must determine if that factor could affect the outcome separately from the intervention." Ibid., 20.

73. Agodini et al., *Achievement Effects of Four Early Elementary School Math Curricula.*

74. Cabalo and Vu, *Comparative Effectiveness of Carnegie Learning's Cognitive Tutor Algebra I Curriculum.*

75. Ibid., 1.

76. Several states not currently in the Common Core have also recently changed their mathematics standards to ensure their students are "college and career ready." Textbook publishers will also have to provide revised editions to these states if they are to align with their more rigorous standards.

77. The eight Common Core mathematical standards are: make sense of problems and persevere in solving them; reason abstractly and quantitatively; construct viable arguments and critique the reasoning of others; model with mathematics; use appropriate tools strategically; attend to precision; look for and make use of structure; and look for and express regularity in repeated reasoning.

78. Lisbeth Schorr and Anthony Bryk, "To Achieve Big Results from Social Policy, Add This," Huffington Post, January 12, 2015, http://www.huffingtonpost.com/lisbeth-lee-schorr/to-achieve-big-results-fr_b_6510262.html.

79. Jennifer O'Day and Marshall S. Smith, "Quality and Equality in American Education: Systemic Problems, Systemic Solutions," in *The Dynamics of Opportunity in America*, ed. I. Kirsch and H. Braun, February 2016, 299–360.

80. Another major change facilitated by the Internet and the nation's move to the Common Core is the increasing use of open curricula, such as the Engage NY mathematics curriculum used by much of New York State and many other districts in the country. Large libraries of Open Educational Resources provide free materials that supplement or replace traditional textbooks. In addition, many individual teachers now modify their curricula using open materials such as online video, simulations, and games. Technology will continue to make curriculum games, simulations, and project-based learning more interesting and pedagogically useful as students become adept in using these powerful tools. US Department of Education, "High-Quality and Easy-To-Use Resources Draw Educators from Around the Nation to EngageNY," Progress: Teachers, Leaders and Students Transforming Education, http://www.ed.gov/edblogs/progress/2014/11high-quality-and-easy-to-use-resources-draw-educators-from-around-the-nation-to-engageny/.

81. In an essay on RCTs, Nancy Cartwright discussed the possibility of design and implementation flaws and argued for engaging experts with "subject-specific knowledge" in areas such as health or education to reduce threats to the internal validity of the study. "It is important though that these are not people like me (or independent experimental-design firms) who know only about

methodology, but rather people with subject-specific knowledge who can spot relevant differences that come up. But this introduces expert judgment into the assessment of internal validity, which RCT advocates tend to despise. Without expert judgment, however, the claims that the requisite assumptions for the RCT to be internally valid are met depend on fallible mechanical procedures. Expert judgments are naturally fallible too, but to rely on mechanics without experts to watch for where failures occur makes the entire proceeding unnecessarily dicey." Nancy Cartwright, *Are RCTs the Gold Standard?* Centre for Philosophy of Natural and Social Science Contingency and Dissent in Science, 2007, http://www.lse.ac.uk/CPNSS/research/concludedResearchProjects/ContingencyDissentInScience/DP/Cartwright.pdf.