Article

American Journal of Evaluation
2017, Vol. 38(4) 471-492

DOI: 10.1177/1098214016662338
journals.sagepub.com/home/aje
SAGE

The Threshold and Inclusive Approaches to Determining "Best Available Evidence": An Empirical Analysis

Jean Stockard[1,2] and Timothy W. Wood[2]

Abstract

Most evaluators have embraced the goal of evidence-based practice (EBP). Yet, many have criticized EBP review systems that prioritize randomized control trials and use various criteria to limit the studies examined. They suggest this could produce policy recommendations based on small, unrepresentative segments of the literature and recommend a more traditional, inclusive approach. This article reports two empirical studies assessing this criticism, focusing on the What Works Clearinghouse (WWC). An examination of outcomes of 252 WWC reports on literacy interventions found that 6% or fewer of the available studies were selected for review. Half of all intervention reports were based on only one study of a program. Data from 131 studies of a reading curriculum were used to compare conclusions using WWC procedures and more inclusive procedures. Effect estimates from the inclusive approach were more precise and closer to those of other reviews. Implications are discussed.

Keywords

best evidence reviews, cumulative science, evidence-based practice, What Works Clearinghouse, Direct Instruction, *Reading Mastery*

Evidence-based practice (EBP), the notion that medical, social, and educational procedures should be based on strong scientific research, emerged in the 1990s. The goal was applauded by social researchers and evaluators, for it seemed to embody the ultimate aim of much of their work—to promote the most effective and equitable policies and programs. Within a relatively short period of time, government bodies embraced EBP and encouraged its application in areas ranging from medicine to education. A variety of groups emerged to summarize the "best available evidence," the most noted of which are perhaps the Cochrane Collaboration, which examines health-care interventions; the Campbell Collaboration, which looks at social, behavioral, and educational areas;

[1] Department of Planning, Public Policy, and Management, University of Oregon, Eugene, OR, USA
[2] National Institute for Direct Instruction, Eugene, OR, USA

Corresponding Author:
Jean Stockard, Department of Planning, Public Policy, and Management, University of Oregon, Eugene, OR 97403, USA.
Email: jeans@uoregon.edu

and the What Works Clearinghouse (WWC), which looks at educational programs. The groups typically use a ''threshold,'' rule-based approach to examining studies, excluding those that do not meet a set list of criteria including, most prominently, a preference for randomized control trials (RCTs). Thus, they assume, at least implicitly, that the most valid results of evaluations emerge from studies with these designs and, often, other characteristics.

While the general goal of EBP has been widely embraced by evaluators throughout the world, the nature of the threshold approach, including its strong preference for RCTs, has been questioned. For instance, in 2003, the U.S. Department of Education solicited comments on a proposal to prioritize evaluation plans that utilized RCTs and, under certain conditions, quasi-experimental designs. The American Evaluation Association (AEA) submitted a statement that opposed the priority, contending that it represented a misunderstanding of methodologies that were scientifically rigorous and the types of studies that could determine causality and legitimately inform policy and program decisions. The AEA, and others, including the American Educational Research Association and the National Education Association, advocated a more inclusive approach to developing research summaries. They voiced concerns that the proposed priority would result in policy recommendations based on only part of the available literature, that some programs might remain unevaluated, and that the public could be deprived of a full and representative understanding of the research findings (AEA, 2003). The Department of Education received close to 300 comments on their proposed priorities, and over 90% expressed concerns such as these. However, despite the objections, the proposed policies were formally adopted with only a small editorial correction (Office of the Federal Register, National Archives and Records Administration, 2005).

As recommendations for best practices based on the threshold type of summary have appeared, a number of authors have expressed concerns about the review process and the validity and/or usefulness of the results. These criticisms have appeared within education (e.g., Confrey, 2006; Hempenstall, 2014; Oancea & Pring, 2008; Shoenfeld, 2006; Slocum, Detrich, & Spencer, 2012; Slocum, Spencer, & Detrich, 2012), medicine (Drake, Latimer, Leff, McHugo, & Burns, 2004), psychotherapy (Kazdin, 2004), and public health (Schmidt, 2014). At the same time, others have continued to defend the threshold-based approach quite staunchly. For instance, some members of the AEA submitted a document, sometimes termed the ''Not AEA Statement,'' supporting the Department of Education's priorities. Conferences, books, and articles have highlighted the seemingly quite contentious debates (Donaldson & Christie, 2005; Donaldson, Christie, & Mark, 2009; Donaldson, Patton, Fetterman, & Scriven, 2010; Gargani & Donaldson, 2011).

To date, discussions of the relative merits of the threshold and inclusive approaches to developing evidence appear to involve largely theoretical or philosophical discussions as well as specific critiques of findings or methods. We suggest, however, that the debate regarding evidence for best practices can be seen as an empirical question, one that can be addressed by examining data. To what extent do research summaries using the threshold method reflect the available base of literature? To what extent does the application of the threshold-type rules affect the estimates of effects that are obtained and the summaries of evidence that would be given to the public?

This article examines these issues using WWC reports and procedures regarding literacy programs. We begin with background literature, contrasting the WWC's rule-based, threshold methodology with the traditional social science literature. We then present results of two empirical studies. The first is descriptive in nature, examining the outcomes of 252 WWC reviews of literacy programs. We look at the probability that research studies would pass the WWC thresholds and the extent to which summary reviews might reflect the range of research available. The second focuses on 131 evaluations of one literacy program. It compares the results that would be obtained using the WWC threshold procedures with those using a more inclusive approach. The final section summarizes the findings and discusses some of their implications.

The WWC Threshold Approach and Traditional Methodological Literature

One could suggest that the threshold approach to finding best evidence reflects, at least implicitly, the writings of the British statistician R. A. Fisher (1925, 1935) that showed how the logic used to develop experiments within agriculture, with complete control over experimental conditions, could be applied to other areas of inquiry. The more inclusive approach to research design and summaries embodies the writings of Donald Campbell and a series of coauthors. We refer to this series of books as the Campbell, Cook, Shadish, and Stanley (CCSS) tradition, reflecting the four authors involved in the key publications (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). They showed how the logic of experimental design can be much more flexible than implied by Fisher. The CCSS tradition describes a wide variety of research designs for field settings that are internally valid. The authors also discuss the importance of external validity or generalizability and stress the cumulative nature of science and the importance of systematically contrasting the results of multiple tests across varying samples, settings, and outcome measures. The CCSS writings have become the standard reference to research design in the social sciences and evaluation research (Chen, Donaldson, & Mark, 2011).

The focus on ''scientifically based research'' in the U.S. Department of Education can be traced to language within the No Child Left Behind Act of 2001. This was shortly followed by a 2002 National Research Council (NRC) report designed to ''review and synthesize recent literature on the science and practice of scientific educational research'' (p. 1). The wording in both documents reflects the writings of the CCSS tradition, emphasizing the importance of flexibility in research design and choosing methodologies that are appropriate for a given question and context. The NRC report also stresses the importance of replicating results and testing generalizations of findings with multiple methods and in a wide variety of settings, mirroring the long-established notion of a cumulative science (Popper, 1962) and the CCSS discussions of generalized causal inference (Shadish et al., 2002).

The departure from this inclusive, broad, and cumulative vision and the emphasis on a more exclusive, threshold-based approach to reviews seems to have appeared after the establishment of the WWC in 2002. The initial statement of the priorities appeared in the Federal Register in 2003 and was formally adopted in 2005.

The WWC's (2014) stated mission is to develop summaries of the ''scientific evidence for what works in education'' (p. 1). WWC reviews are conducted by specialized teams consisting of a content expert, methodological expert, and review staff and are guided by the WWC *Procedures and Standards Handbook*, which details the elements involved in the review process. The first edition of the handbook and the first summary reports appeared in 2008. The handbook has been revised multiple times since its inception to provide greater clarification on active WWC policies and add new elements to the review process. For instance, while the original Federal Register notice involved only the discussion of acceptable or preferred research designs, later versions of the handbook added a variety of other conditions to determine whether a given study provided credible evidence for best practices and should be included in a review.

The WWC review process begins by developing a protocol defining the topic and scope of the analysis and identifying a group of potentially relevant studies. At this point, the review team examines each identified study to see if it is eligible for review, the step we refer to as Threshold 1 (T1). Studies that pass T1 are then examined to see if they meet certain ''standards of evidence,'' a stage we refer to as Threshold 2 (T2). The following sections examine these thresholds using the CCSS writings as a touchstone.

T1 Criteria

At T1, the WWC examines the identified studies to ensure that they meet five general criteria. The first limits the set of studies reviewed to those that match the topic of the analysis, omitting a study if it is "not a primary analysis of an intervention," such as a meta-analysis or literature review (WWC, 2014, p. 7). The second criterion calls for omission if a study "does not include an outcome within a domain specified in the protocol" (WWC, 2014, p. 8). For example, with studies of reading, those that assess specific skills, such as reading fluency, comprehension, or general reading ability, would be accepted while those that involve measures such as high school grades, scores on general ability tests, or student self-confidence or behavior would be eliminated. The third criterion calls for the omission of studies that do not involve a defined "age, grade range, gender, or geographic location." Separate reviews of reading curricula are developed for general education students, students with learning disabilities, those in different grade levels, and English language learners. The second and third criteria help produce a review set that is relatively homogeneous. Yet, the CCSS tradition notes that variability within a group of studies is a key element in determining external validity and estimates of replicability. Such generalizations can only be made when one examines "the extent to which the causal relationship holds over variation in persons, settings, treatment, and measurement variables" (Shadish et al., 2002, p. 341).

The fourth T1 criterion calls for omitting studies that were "not published in the relevant time frame ... usually 20 years prior to the start of the WWC review" (WWC, 2014, p. 8). In its analysis of educational research, the NRC committee emphasized the importance of time, noting that the accumulation of knowledge within a field can take decades or even centuries (NRC, 2002, p. 44) and that knowledge can be constantly evolving, as new theories, methods, and analysis techniques develop. Thus, an inclusive approach would also, presumably, not limit a set of review studies by date of publication but would instead examine the relationship of date of publication to estimates of effects.

The fifth T1 criterion concerns research design, restricting analyses to pretest–posttest control group designs. Studies that randomly assign subjects to groups can be accepted "without reservations," while those that use other means of group assignment can only be accepted "with reservations" (WWC, 2014, pp. 7–10).[1] While the CCSS writings describe the importance of randomized experiments in establishing "causal description," the authors are also quite clear in stressing the importance of using other types of designs, noting that "experiments are far from perfect means of investigating causes" (Shadish et al., 2002, p. 8):

> Among scientists, belief in the experiment as the *only* means to settle disputes about causation is gone, though it is still the preferred method in many circumstances. Gone, too, is the belief that the power experimental methods often displayed in the laboratory would transfer easily to applications in field settings. (p. 30, emphasis in original)

For instance, it is often very difficult to conduct randomized studies involving individuals within field settings, such as schools, in which the participants are blinded to the condition of treatment. Teachers, and often students, are well aware of how their experiences differ from peers. This knowledge, as well as communications between members of treatment groups, would, by definition, threaten internal validity of the design. Similarly, it is often very difficult in field settings to control other issues that may affect internal validity, such as attrition. Given these realities, the CCSS tradition proposes numerous alternatives, each of which is seen as having internal validity that is equal to or greater than that of RCTs in such situations.

Three of these alternatives, none of which involves random assignment, could, potentially, be especially useful in educational research. The first is "recurrent institutional cycle" or "cohort control group" designs, described as useful in organizational settings, such as schools, that have

a regular turnover of cohorts or "graduates," as long as the groups differ in only minor ways and the intervention is or is not given to all group members (Campbell & Stanley, 1963, pp. 56–61; also Cook & Campbell, 1979, pp. 126–127; see Shadish et al., 2002, pp. 148–149). A second is "normed comparison contrasts," in which changes over time for a treatment group are compared to published norms (Shadish et al., 2002, pp. 126–127). Norm-referenced designs have been found to yield estimates of student gains that are comparable to those obtained in RCTs and were once promoted by the U.S. Department of Education (Tallmadge, 1977, 1982). A third alternative is interrupted time series, described by Shadish and associates as "one of the most effective and powerful of all quasi-experimental designs" (Shadish et al., 2002, p. 171). Several authors have described how a multiple baseline interrupted time-series approach may be especially well suited to long-term evaluations of large group interventions, such as those occurring in education, and produce results that are extremely close to those resulting from RCTs (Biglan, Ary, & Wagenaar, 2000; Biglan, Flay, Komro, Wagenaar, & Kjelistrand, 2012; St. Clair, Cook, & Halberg, 2014).

Notably, of these three designs, only the norm comparison design requires pretesting of subjects. The CCSS definition of randomized experiments does not mention pretests, focusing instead on the assignment of units to conditions and posttest assessment. Four of the nine examples of randomized designs listed in a summary table in the latest CCSS volume do not include pretests (Shadish et al., 2002, p. 258). While noting that issues of group equivalence become more complex when random assignment is not used, they also provide numerous examples of internally valid quasi-experimental designs that incorporate control groups without a pretest of all subjects (see Shadish et al., 2002, pp. 115–130).

T2 Standards of Evidence and the CCSS Methodological Tradition

Studies that pass T1 are then examined to see if they meet various standards of evidence, none of which appears to have been anticipated in the Federal Register posting of 2003 or the 2005 final determination. The first standard commonly used to exclude studies involves pretest equivalence of treatment and control groups.[2] It requires that groups differ by no more than .25 of a standard deviation (*SD*) on all baseline (pretest) characteristics. If the difference falls between .05 and .25 of an *SD*, statistical adjustments must be used (WWC, 2014, p. 15).[3] The requirement applies to studies that have employed random assignment or other means of assigning subjects to groups, such as matching. It excludes the possibility that differences at pretest might go in opposite directions on various measures (e.g., an intervention group being higher on one measure and lower on another), that the average of multiple differences at pretest falls below the threshold, or that the magnitude of the treatment effect might surpass the magnitude of any pretest difference. Yet, differences larger than the WWC criteria are likely to occur by chance, especially with multiple outcome measures or with smaller samples, as often occurs with randomized studies. The CCSS literature discusses this possibility, noting that random differences in pretest characteristics can influence the results of a study and again emphasizing the importance of multiple examinations of research questions (Shadish et al., 2002, p. 250).

The other most commonly used T2 standards involve potential confounds or threats to internal validity. One requires that the data for the treatment and comparison group come from the same year (WWC, 2014, p. 20), a standard that automatically excludes studies using cohort control group designs and many time-series designs. Another, termed "one unit per condition," excludes studies "when all of the intervention students are taught by one teacher, all of the comparison classrooms are from one school, or all of the intervention group schools are from a single school district" (WWC, 2014, p. 19). Yet, studies that involve comparisons between two schools or two districts often have data for numerous groups within each setting. Multilevel modeling techniques can control for such factors and estimate their impact. Studies with only one teacher have tried to control

for this issue by methods such as having that person use both interventions. In addition, the requirement of multiple schools within a district could automatically eliminate small rural districts from consideration and require very large samples. Thus, adherents of the CCSS tradition could question automatic exclusion of studies using this standard.

Two characteristics of studies that could potentially impact estimates of effects are not included within the WWC standards: fidelity of implementation and dosage. The WWC (2014) explicitly notes that it does not exclude studies with low levels of fidelity (p. 21) and does not appear to consider dosage at either T1 or T2. Both of these areas are, however, generally considered a key element of internal validity or ''confounding'' and can influence estimates of the efficacy of an intervention. If a program were effective, greater dosage should logically result in stronger improvement. If a program were ineffective or harmful, greater dosage would result in no change or even decline. Poor implementation of an ineffective program could result in assessments that provided overly positive reports, while poor implementation of an effective program could result in assessments that provided overly negative reports (Stockard, 2010; see also Confrey, 2006; Shadish et al., 2002, pp. 315–321; Zvoch, 2012).

Ratings of Effectiveness

Using the studies that pass T2, the WWC develops summary ratings of effectiveness. The ratings are based on the average effect size across all subgroups and outcomes included in a particular study and a count of the number of positive and negative effects within the group of accepted studies. An intervention can receive a rating of ''positive effects'' if there are two or more accepted studies that ''show statistically significant positive effects, at least one of which meets WWC group design standards without reservations, AND no studies show statistically significant or substantively important negative effects'' (WWC, 2014, p. 29, emphasis in original). A rating of negative effects is simply the opposite. Statistical significance is defined as the .05 level and substantive significance as an effect size of absolute value .25 or larger. A ''potentially positive'' or ''potentially negative'' rating is given when at least one study shows significant positive effects, fewer or the same number of studies have indeterminate effects, and no studies have negative effects. A ''mixed'' rating is given when there are both significant positive effects and significant negative effects. Thus, a set of studies that included one with negative results, but many with positive results, could receive only a mixed rating.

From the perspective of a more inclusive methodological tradition, these requirements could raise a number of questions. One involves the requirement that all study results be positive for a positive rating (or negative for a negative rating). When there are more studies within a field, there is a greater probability that some will have opposite findings. That is, interventions that have a larger base of efficacy studies would be more likely, simply by chance, to have at least one significant contrary result. This issue could be compounded by the failure to consider size of the study sample. For example, a set of studies with a negative result based on only a few students, but positive results from many more studies involving larger samples, would still result in a mixed rating. In addition, the aggregation of results across studies ignores possible variations in the findings, which might result from different outcomes related to measures, subgroups, or design elements, including fidelity and dosage. Such variations are important for knowing the extent to which results are consistent across settings, a key element of external validity. Finally, those using a more inclusive approach could question the fact that only two studies are needed to make summary judgments.

Summary

The material reviewed above indicates that the exclusive, rule-based WWC approach to developing best evidence reviews differs in a number of ways from traditional social science

methodology. The T1 screening criteria could result in the selection of a narrow and homogeneous set of studies, thus making assessment of external validity more difficult. These criteria also result in the exclusion of studies that use designs traditionally seen as internally valid and especially appropriate for field settings. Standards applied at T2 may further narrow the range of studies examined, with those involving multiple outcome measures as well as smaller samples and rural districts perhaps particularly affected. At the same time, the T2 standards do not include measures of dosage and fidelity, typically seen as potentially important influences on program effectiveness. Finally, the WWC's method of developing summary ratings could potentially result in misleading conclusions regarding large bodies of literature and mask variations in results that would be an important element in establishing external validity.

We turn now to empirical analyses of the results of the WWC procedures designed to examine the extent to which these theoretical concerns are reflected within data. The first study examines the extent to which the procedures affect the proportion of research results included in reviews. The second study uses data from 131 studies of one intervention to examine the way in which each of the WWC criteria and standards, as well as other factors not used by the WWC, affect estimates of the impact of the program.

Study 1: A Descriptive Analysis of WWC Conclusions

The responses to the 2003 Federal Register statement of priorities and the literature reviewed above suggest that the WWC criteria and standards could limit the material used to present summary evaluations. Our first study was designed to assess this possibility.

Methodology

We examined WWC reviews of literacy-related interventions from 2008 to mid-2014. The WWC website indicated that a total of 252 intervention programs had been examined: the majority under the Beginning Reading Protocol ($n = 175$), and the remainder regarding Adolescent Literacy ($n = 25$), Early Childhood Education ($n = 15$), English-language Learners ($n = 30$), and Students with Learning Disabilities ($n = 7$). For each intervention report, we noted the year of publication, the number of studies examined, the number of studies that passed T1, the number of studies that met T2 with and without reservation, the WWC's summary judgment of the intervention, and, for interventions that had studies passing T2, the total number of students in these studies. We also calculated the probability that a study would meet each threshold. The WWC generally did not report the total number of studies examined if none passed T1 nor the reasons that studies were not included. In addition, after 2008, it appears to have only reported on interventions that had at least one study that passed T1. Thus, at least some of our measures are underestimates of the probability that a study or intervention would be included for review.[4]

Our analysis was descriptive in nature. We examined results for the total sample as well as, for the T2 decisions, variations by year of publication (distinguishing those published in 2009 or earlier, 44% of the sample, from those published more recently), the subject area protocol used for the report, and the final WWC judgment (distinguishing those with mixed, no effects and/or some negative judgments, 42% of the sample, from others). Some of the intervention reports had a substantially larger research base than others. Thus, both medians and means are reported. Results are given for the full sample and for a reduced sample that omitted six interventions with a total number of studies that was more than two *SD*s above the mean.

Table 1. WWC Judgments of Studies of Literacy Programs at Threshold 2, July 2007–August 2014.

Number and Proportion of Studies Accepted for Review				
	All Programs ($n = 93$)		Outliers Removed ($n = 87$)	
Group	*N*	Proportion	*N*	Proportion
Total studies identified	4098	—	2453	—
Met T1 criteria	436	0.11	307	0.13
Met T2 standards without reservation	93	0.02	78	0.03
Met T2 standards with reservation	69	0.02	61	0.03
Total met T2 and included in review	162	0.04	139	0.06
Average Number of Studies Accepted at Each Tier and Number of Students in Studies Used in WWC Reviews				
	All Programs ($n = 93$)		Outliers Removed ($n = 87$)	
Group	Mean	Median	Mean	Median
Studies identified	44.1	16	28	14
Studies that passed T1	4.8	2	3.5	2
Studies passed T2 without reservations	1	1	0.9	1
Studies passed T2 with reservations	0.7	0	0.7	0
Total met T2 and included in review	1.7	1	1.66	1
Number of students in studies reviewed	776	146	753	140

Note. T1 = Threshold 1; T2 = Threshold 2; WWC = What Works Clearinghouse. Results are given for the full sample of programs with at least one study accepted at T1 (93 programs) and for a reduced sample that omitted six interventions with a total number of studies that was more than two standard deviations above the mean (87 programs). The WWC was unable to determine the number of students in three studies.

Results

Only a minority of the 252 intervention programs (37%, $n = 93$) had at least one study that met the T1 criteria and were assessed at T2. Table 1 summarizes WWC judgments of these 93 interventions at T2. The top panel reports the number and proportion of studies that met each threshold. The data indicate that slightly more than one tenth of the studies that were examined for these interventions passed T1 and less than half that proportion passed T2 and were considered in summary reports. Similar patterns appeared in the various subject areas reviewed, the year in which the reports were published, and for those with different summary judgments.[5]

The bottom panel of Table 1 provides insight into the basis for an "average" intervention report. The first five lines report the average number of studies identified and retained at each stage of the review, and results parallel those shown in the top panel. For instance, looking at median values for the sample with outliers removed (the far-right column, $n = 87$ interventions), it can be seen that half of the programs had at least 14 studies identified for review. Yet, on average, for a given intervention, two or fewer studies passed T1 and only one study passed T2. In total, 75 interventions (81% of the 93 interventions examined at T2 and 30% of the 252 interventions initially identified) were judged to have at least one efficacy study worthy of a summative judgment. Almost half of the published summary reports (36/75 or 48%) were based on one study, over a fourth were based on only two studies (20/75 or 27%), and only six were based on more than four studies.

Given how few studies were included in the reports, it is perhaps not surprising that half of the reports relied on data from 140 students or less or that the reports did not report definitive conclusions. There were no interventions that received a positive or negative rating on all the dimensions examined, although almost three quarters of the reports (54/75) cited positive or potentially positive

results on some dimensions. The remainder reported no discernible or mixed effects (16/75) or some negative effects (5/75).

Summary

As expected, this examination of WWC judgments of 252 intervention programs revealed that the review process resulted in a substantial narrowing of the body of research eligible for summary within a review. As a result, the majority of the best evidence reports were based on only one or two research studies, involving fewer than 200 students. Scholars working within the CCSS methodological tradition could question the extent to which summaries based on so few studies could accurately reflect a cumulative research literature and suggest that a more inclusive analysis would be more informative. Our second research question examined this issue.

Study 2: Comparing the Threshold and Inclusive Approaches to Research Summaries

Study 2 was designed to systematically compare the conclusions that would result from using the WWC's threshold approach and a more inclusive analysis. To this end, we examined a relatively large body of literature regarding one intervention. We addressed three general questions: (1) How likely is it that research studies would pass T1 and T2 and which criteria and standards are most likely to influence acceptance of a study? (2) How are estimates of effect size associated with each of the criteria and standards as well as with other characteristics of the studies? (3) How would summary judgments of effectiveness using the WWC procedure and a more inclusive method differ?

Methodology

To minimize issues associated with small samples and heterogeneous results, we chose to examine an intervention with a well-established literature base and relatively consistent findings: *Reading Mastery*, an instructional program that is part of the Direct Instruction (DI) corpus of curricula. In a meta-analysis that compared 29 different programs, Borman and associates found the strongest results for DI and noted that "the research base for Direct Instruction (DI) is very extensive and of very good quality" (Borman, Hewes, Overman, & Brown, 2003, p. 187; see also Adams & Engelmann, 1996; Coughlin, 2014; Hattie, 2009). Using published bibliographies and searching a variety of databases, we identified 131 efficacy studies that used a design that allowed comparisons of achievement between students who were or were not exposed to the program. To maximize sample size, we included studies about other versions of the program but tested for any differences by program title in our analysis.[6]

Online Appendix A has a full list of studies included in the analysis and key variables associated with each. Online Appendix B has supplementary methodological details on measures, calculations of effect size, and statistical results omitted from the text to conserve space.

Measures. For each research report, we coded information related to the criteria and standards involved in T1 (nature of the design, literacy area studied, grade level of the sample, and year of publication) and T2 (extent of pretest differences and presence of potential confounds). We also coded a number of other characteristics of the studies that could be of interest to consumers of best evidence reviews including (1) areas that would generally be used in tests of external validity, such as students' race–ethnicity, poverty status, special education status, and schools' location and level; (2) the type of assessment measure used (e.g., norm based, curriculum based, state assessment, or developed by the researcher); (3) dosage; (4) teachers' experience with the program; (5) whether or not the data were gathered after the intervention ended (i.e., maintenance measures); (6) indicators

of fidelity of implementation and teacher training; and (7) the number of students involved in each comparison. The sample of research studies reflected substantial diversity. They used different efficacy measures, different research designs, samples of students from all regions of the nation as well as other countries, rural and urban areas, special education and general education students, relatively short periods of exposure to the program as well as longer periods of exposure, different levels of implementation fidelity, and immediate as well as long-term outcomes (see Online Appendix B, Table B-3).

Effect sizes were calculated for each reported outcome using Cohen's *d*, calculated as the difference between the posttest means of a treatment and control group divided by the common *SD*. The studies varied in the types of posttest data that were reported, such as including percentages or results of inferential statistics or regression equations rather than means or *SD*s. We used a variety of methods to convert these alternative data to Cohen's *d* and examined differences between these calculation methods in our statistical analyses.[7]

Outliers, defined as effects more than 3.0 *SD*s above or below the mean and comprising less than 1% of the total number of effects, were omitted. We also omitted effect sizes that were aggregates of others in the analysis.[8] For instance, if a study reported scores for individual reading domains, such as comprehension and vocabulary, as well as an aggregate of these two domains, we omitted the aggregate score. Such aggregates comprised 17% of the total number of effects that were calculated. In total, 1,353 effect sizes were included in the statistical analyses.

Analysis. To address the first question, we examined the percentage of studies that would pass each threshold and be included in a WWC report. Given the results of Study 1, we expected that relatively few studies would pass both of the thresholds.

To address the second research question, we used mixed models with effect sizes as the dependent measure and study design as a random effect. This approach was chosen because each of the studies in our analysis could have multiple outcome measures and sometimes employed multiple designs. Because characteristics of design are so central to the exclusion decisions in T1 and T2, we chose to use design, rather than study, as the random effect. However, results using study as the random effect were substantively identical to those reported below and are given in Online Appendix B.[9] Variables measured at the effect level were treated as Level 1 variables, while characteristics of the design were treated as Level 2 variables. We began with a baseline, intercept-only model. This is equivalent to an analysis of variance with design as the independent variable. The coefficient associated with the intercept is equivalent to an estimate of the average effect size across all designs. We then examined the relationship of the T1 and T2 criteria and standards, as well as other study characteristics, to the estimates of effects. We examined the differences from the baseline model using the -2 log likelihood measure of model fit and the magnitude and significance of the coefficient associated with the intercept. The impact of each criteria, standard, and study characteristic was examined singly as well as jointly. Reduced summary models, which include all variables found to be significant in the other analyses, are given below. Results of the intermediate models are in Online Appendix B Tables B-4, B-5, B-6, and B-8.

If the criteria and standards used in the thresholds were important in providing more accurate estimates of effects, we would expect that changes in model fit and the associated coefficients would be significant. Alternatively, the logic of the CCSS tradition would lead one to expect that, if a program were highly effective, few of the defined criteria and standards associated with the thresholds would have a significant relationship. That is, estimates of effects would be robust to variations in design and threshold related variables, and similar results would appear across the various analyses.

Based on the consistent results in other studies of the curriculum, we also expected that there would be few, if any, significant influences of the way in which the effect size was calculated, characteristics of the students and their schools, or the specific DI reading program that was used. In

Table 2. Designs Employed in the Studies.

Design	*N*	Percentage (Studies)	Percentage (Designs)
Pretest–posttest designs, accepted with no reservations			
Pretest–posttest control group with random assignment design	24	18.3	16.2
Pretest–posttest designs accepted with reservations			
Pretest–posttest control group design	35	26.7	23.6
Pretest–posttest control group with statistical controls design	8	6.1	5.4
Pretest–posttest matched control group design	4	3.1	2.7
Pretest–posttest cohort control group design	5	3.8	3.4
Pretest–posttest designs, not accepted by What Works Clearinghouse			
Pretest–posttest norm comparison design	7	5.3	4.7
Cohort control group historical comparison design	15	11.5	10.1
Cohort control group norm comparison design	10	7.6	6.8
Pretest–posttest gain score design	6	4.6	4.1
Multiple baseline time-series design	1	0.8	0.7
Posttest-only designs			
Posttest-only control group design	13	9.9	8.8
Posttest-only matched control group design	1	0.8	0.7
Posttest-only norm comparison design	2	1.5	1.4
Cohort control group design	17	13.0	11.5
Total	148	113.0	100.0

Note. One hundred thirty-one studies with a total of 148 designs were included in the analysis. Ten of the 131 studies in the analysis reported results with two different designs and one reported results with three designs. The percentages associated with the studies were calculated using the number of studies as a base ($N = 131$) and thus do not add to 100.

contrast, given the previous findings, we expected that indicators of both dosage and fidelity would influence the results, with greater dosage and higher fidelity having a positive impact. In addition, we expected that postimplementation (maintenance) estimates of effects would be positive, although smaller than those found immediately after an intervention ended.

To answer the third research question regarding summary judgments, we followed the WWC guidelines for calculating the average effect size and statistical significance for each study and counted the number of positive and negative results. We looked at the results for all 131 studies, those that passed T1, and those that were close to passing T2. (As shown below no studies would pass T2.) We then compared these results to averages and confidence intervals (CIs) of the effects within each group.

Results

Research Question 1: How many studies would pass the WWC thresholds? Table 2 summarizes the designs used in the studies. Less than a fifth of the studies used the WWC preferred pretest–posttest control group design with random assignment, although an additional 40% used a design that could be accepted at T1 with reservation. Almost a third used a pretest–posttest control group design not accepted by the WWC but recommended by the CCSS tradition, such as norm comparison, cohort control groups with historical comparisons, gain score, or time-series designs. A quarter used some type of posttest-only control group design.

Table 3 reports the number of studies that would pass T1 and T2. Most of the studies met the T1 criteria (top panel) regarding domain and required grade range. The WWC protocols for reading use two different dates of publication for exclusion at T1 (1983 and 1989), and the majority of the

Table 3. Number and Percentage of RM Studies Passing T1 and T2.

Number and Percentage of Studies Passing T1		
T1 Criteria	*N*	%
Measures a specified reading domain	123	93.9
Grade level		
K to Grade 3	62	47.7
Grades 4–12	24	18.2
Grades K–12	37	28.0
Pre-K or Pre-K to K or Grade 1	8	6.1
Year of publication		
1983 or later	110	84.1
1989 or later	98	74.1
Design		
Design acceptable without reservation	24	18.3
Design acceptable with reservation	48	36.6
Met all T1 criteria, published 1983 or later		
Without reservation	12	9.2
With reservation	34	26.0
Met all T1 criteria, published 1989 or later		
Without reservation	10	7.6
With reservation	29	22.1

Number of Studies Passing T2, by Year Criteria		
T2 Standards	Published 1983 or later ($n = 46$)	Published 1989 or later ($n = 39$)
Standards regarding pretests		
Had pretest data	32	30
No pretest differences greater than .25 *SD* in absolute value	7	6
No pretest differences greater than .05 *SD* in absolute value	2	2
Standards regarding confounds		
One-unit rule—more than one district per condition	7	7
One-unit rule—more than one school per condition	35	29
Data for each condition from the same year	42	35
No other DI reading program used	37	30
Met all but one standard of evidence ($\leq$.25 *SD* difference)	6	5
Met all but one standard of evidence ($\leq$.05 *SD* difference)	2	2
Met all standards of evidence	0	0

Note. $N = 131$. RM = Reading Mastery; DI = Direct Instruction; T1 = Threshold 1; T2 = Threshold 2.

studies were published in 1983 or later. Over half of the studies met each of the separate criteria at T1 (reading domain, grade level, year of publication, and design), but only about one third met all of the specified criteria and would thus be eligible for screening at T2. Less than 10% of the studies met all the T1 criteria without reservation.

The results in the second panel of Table 3 are restricted to studies that passed T1. Results are given separately for those meeting the two criteria regarding date of publication. Close to three fourths of the studies reported data from which pretest differences could be calculated,[10] but fewer than 10 studies had pretest differences that met the standard of .25 *SD* difference on all measures, and only 2 had all pretest differences smaller than .05 *SD*.[11] A majority of the studies would meet the standards regarding confounds related to data from only one school, from different years, or from

using multiple DI programs. Yet, only seven studies would meet the standard regarding data from more than one district in each condition. Taken together, none of the studies that passed the first threshold would meet all of the T2 standards of evidence. Six of the studies that passed T1 would meet all but one of the T2 standards if the more liberal 1983 date of publication criterion and .25 *SD* pretest difference standard were used.

Research Question 2: Influences on Effect Sizes. As expected, there were few significant relationships between effect size and the T1 criteria and T2 standards. Of the 17 measures of T1 criteria examined, only the grade level of students was significant. Of the eight T2 standards, only the measure associated with data from one district was significant. There were no significant influences on effect sizes of the type of assessment, method of calculating effects, or other characteristics of students or schools; and results held across the different variations of the reading program. As expected, measures of dosage, maintenance, and fidelity were significant.

Table 4 summarizes results of the reduced mixed model analyses regressing effect sizes on the criteria, standards, and study characteristics that were significant in the individual analyses. The first model includes only the T1- and T2-related measures, the second model includes only study characteristics, and the third includes all of the variables. In the final model, only 1 of the 24 indicators related to the criteria and standards and only 3 of the 21 study characteristics related to measures and student and school characteristics were significantly related ($p = .05$) to the estimates of effect size, a proportion of significant effects close to what would be expected by chance. In contrast, most of the coefficients regarding study dosage, maintenance, and fidelity remained significant, with, as expected, larger effect sizes with greater exposure to the program and stronger fidelity. Effect estimates at follow-up periods were positive but smaller than those immediately after the intervention (as indicated by the negative coefficients). Throughout all the models in Table 4 (as well as in Online Appendix B), the overall estimates of the effect size (the intercept) were statistically significant and exceeded the level of .25 used by the WWC to denote substantive significance (ranging from .39 in the baseline model to .79 in Model 3 with all controls), thus replicating earlier analyses of the program. In other words, including the various criteria and standards used at T1 and T2 did not substantively alter estimates of the effects of the intervention.

Research Question 3: Comparing the threshold and inclusive summaries. The mixed models described immediately above derived estimates of effectiveness by using individual effects as the unit of analysis and controlling for design and study related variables. In contrast, Table 5 summarizes estimates of effectiveness with averages and standard errors (used to compute CIs) and the categorization and counting method employed by the WWC. Results are given for the total group of 131 studies, for the subgroups that met T1 with varying criteria, and for the group that came within one standard of meeting T2. These subgroups ranged in size from 6 to 34 studies. (Recall that no studies passed T2 and thus the WWC would not have reached this point in their analysis of effects.)

The average effects, shown in the top panel, ranged from .18 to .62, a difference of .44 *SD* units. As would be expected the standard errors varied substantially, with the smallest values and thus the narrowest CIs, for the larger groups of studies (Groups a, b, and c). All but the results for the smallest subgroup (Group f) were statistically significant.

The second panel of Table 5 reports the summary ratings that would be given for the set of studies within each subgroup using the WWC's system of counting positive and negative effects. In all groups, only a few of the average ratings met the level for being termed a significant negative effect. As would be expected, the larger groups were more likely to have negative results. Because the WWC summary rating of effectiveness is based on comparing the number of ratings in each category, the results with the three largest groups (a, b, and c) would be termed mixed, even though they had the smallest standard errors and, for the full group of 131 studies, the number of studies with positive effects was far

Table 4. Mixed Models, Regressing Effect Size on Selected Criteria, Standards, and Study Characteristics.

	Model 1			Model 2			Model 3		
Variable	Coefficient	*SE*	Probability	Coefficient	*SE*	Probability	Coefficient	*SE*	Probability
T1 criteria									
Grades K–3	0.07	0.04	.05	—	—	—	0.03	0.04	.43
T2 standard									
Data from one district	−0.31	0.07	<.001	—	—	—	−0.16	0.08	.05
Study characteristics									
Effects adjusted for predifferences	—	—	—	0.13	0.06	.02	0.11	0.06	.04
Urban setting	—	—	—	−0.30	0.07	<.001	−0.22	0.08	.01
Rural setting	—	—	—	−0.23	0.08	.01	−0.16	0.09	.06
Log posttest *N*	—	—	—	−0.02	0.01	.17	−0.03	0.02	.05
Dosage, maintenance, and fidelity									
Exposed 3 or more years	—	—	—	0.08	0.04	.04	0.08	0.04	.04
School 3 or more years experience	—	—	—	0.23	0.05	<.001	0.23	0.05	<.001
Follow-up 2 or more years post	—	—	—	−0.30	0.07	<.001	−0.26	0.07	<.001
Lower fidelity	—	—	—	−0.24	0.09	.01	−0.24	0.09	.01
Lower quality study	—	—	—	−0.29	0.12	.01	−0.28	0.11	.01
Constant (adjusted effect size)	0.62	0.08	<.001	0.68	0.10	<.001	0.79	0.12	<.001

Note. Analysis was based on 1,353 effect sizes, 148 designs, 1 to 83 effects per design, with an average of 9.1. The coefficient associated with the intercept in the baseline model was .39, $p < .001$. Each of the models in the table provided a significantly better fit to the data than the baseline model. Variance estimates associated with the random effects for the baseline model indicate that 55% of the variation in effect sizes was between study designs.

larger than the number in the other categories. In contrast, the results with Groups d, e, and f would be termed positive, even though they had the largest standard errors, the associated *t*-ratios were much smaller, the 95% CI for Group f would include zero, substantially fewer studies were included, and the studies represented results from a relatively small number of students.

Summary

In Study 2, we examined numerous replications of efficacy studies of one intervention and the extent to which variations in study design and other characteristics impacted estimates of effects. The findings confirmed expectations based on the literature reviewed in the first part of this article. Even though we examined a literature that has been described as "very extensive and of very good quality" (Borman et al., 2003, p. 187), none of the 131 studies identified would pass all of the WWC criteria and standards and thus be eligible for inclusion in a WWC review. Statistical analyses indicated that the WWC criteria and standards were not related to estimates of effectiveness. Yet, factors omitted from the standards, including implementation fidelity, dosage, and maintenance measures were associated with these

Table 5. Average Effect Sizes, Continuous and Categorical Measures, and Summary Ratings, Total Sample and Subgroups.

Continuous Measures of Effectiveness					
Group	Average Effect	2 × *SE*	Average Sample Size	*t*-Ratio	Total No. of Students
a) Total group	0.41	0.09	647	5.26***	84,749
b) Met T1, 83 rule, with reservations	0.26	0.17	869	3.86***	29,555
c) Met T1, 89 rule, with reservations	0.21	0.20	986	3.35***	28,589
d) Met T1, 83 rule, no reservations	0.53	0.25	55	1.97*	662
e) Met T1, 89 rule, no reservations	0.62	0.25	53	2.28*	534
f) Missed T2 by one standard	0.18	0.25	306	1.54	1,838

Categorical Measures of Effectiveness					
	Number of Effect Sizes by Range				
Group	Negative	Indeterminate	Positive	Total Number of Studies	WWC Summary Rating
a) Total group	6	47	78	131	Mixed
b) Met T1, 83 rule, with reservations	4	16	14	34	Mixed
c) Met T1, 89 rule, with reservations	4	16	9	29	Mixed
d) Met T1, 83 rule, no reservations	0	2	10	12	Positive
e) Met T1, 89 rule, no reservations	0	1	9	10	Positive
f) Missed T2 by one standard	0	4	2	6	Positive

Note. T1 = Threshold 1; T2 = Threshold 2. Average effects were calculated across studies. The column labeled 2 × *SE* (2 times the standard error) can be used to calculate the 95% confidence interval. Average sample size refers to the average number of students in the analyses within each study. The *t*-ratios were calculated using the formula used by the What Works Clearinghouse (WWC): $t = \text{effect} \times \left[\sqrt{(n_i \times n_j)/(n_i + n_j)}\right]$, where n_i and n_j are the average size of the samples in the control and intervention groups. The column "Total No. of Students" is the total number of students across all studies within the group. In the second panel of the table, negative effects are the number of average effect sizes less than or equal to −.25, indeterminate are the number between −.24 and +.24, and positive effects are the number equal to or greater than +.25.
*$p < .05$. ***$p < .001$.

estimates. Finally, the WWC method of providing summary judgments of the literature produced conclusions that differed markedly from those using more conventional and inclusive methods.

Summary and Discussion

It has now been almost two decades since calls for EBP became common and summaries of "best evidence" began to appear. This article has focused on the WWC's approach to summarizing

literature regarding educational interventions. The WWC limits its reports to studies that meet a defined set of criteria and standards, including giving priority to RCTs. When this priority was announced in 2003, a number of individuals and groups, including the AEA, expressed concerns that it could limit the information given to the public and produce misleading results. These concerns reflected the tenets of the classic methodological literature, which advocates using research designs appropriate to a given setting or problem and comparing results across numerous replications. Given the amount of time that has passed since the inception of the WWC it is now possible to empirically examine the nature of their reports and procedures to see the extent to which the expressed concerns materialized. To that end, we conducted two empirical studies. Their results provide insight into the way in which the threshold approach limits the range of material included in best evidence reviews, how conclusions of a threshold approach and a more inclusive approach can vary, and policy changes that could make best evidence reviews more accurate and helpful to decision makers.

The Range of Studies Examined in Threshold and Inclusive Reviews

Our first study examined the results of over 250 WWC reviews of literacy programs. The results appear to confirm the concerns expressed in 2003, for only a small proportion of the identified literature passed the established thresholds. As a result, the WWC summarized data on the efficacy of less than two fifths of the identified interventions. The resulting evidence reviews were most often based on only one study and involved data from relatively few students. None of the summary reports presented a fully positive or negative judgment on an intervention across all the dimensions examined.

Our second study examined how conclusions regarding best evidence would differ using the WWC threshold approach and a more inclusive analysis. We looked at 131 studies of a reading curriculum that had been described as being of high quality and having strong and consistent evidence of effectiveness. All studies used a design identified within the CCSS writings as internally valid. About one third would pass the WWC's first threshold of review, but none would pass the second; and a WWC analysis would conclude that there was no basis for a judgment regarding the program. Thus, the results replicated those of Study 1 by indicating the difficulty in meeting all of the WWC criteria and thresholds. In fact, an even smaller proportion of the studies of the chosen reading intervention met the thresholds than in the analysis of the larger group of literacy programs in Study 1. We are not sure why this result appeared, but suspect that it may reflect our use of the most recent protocol for the threshold approach and a gradual expansion and tightening of the various standards over time, well beyond the description given in the 2003 Federal Register.[12]

Some could suggest that the WWC accepts a very small proportion of the available studies because the general body of research is of such low quality. Prioritizing randomized control designs and applying restrictive criteria and standards help to ensure that only "true experiments" and the best studies are examined. But, a careful examination of the CCSS tradition challenges this view. The concept of true experimental designs was introduced in the first CCSS volume (Campbell & Stanley, 1963) to describe randomized control studies, and it has become a standard component of textbook discussions. Yet, later versions of the CCSS tradition rued its use stating, "We shall not use the term [true experiment] at all given its ambiguity and given that the modifier *true* seems to imply restricted claims to a single correct experimental method" (Shadish et al., 2002, p. 13, emphasis in original). The authors described their work as

> about improving the yield from experiments that take place in complex field settings, both the *quality* of causal inferences they yield and our ability to generalize these inferences to constructs and over variations in persons, settings, treatments, and outcomes. (Shadish et al., 2002, p. 32, emphasis added)

In a later section of this book, they explained that even relatively complex models such as random selection of units from a population followed by random assignment to treatments

> cannot be advocated as *the* model for generalized causal inference .... Though we unambiguously advocate it when it is feasible, we obviously cannot rely on it as an all-purpose theory of generalized causal inference. So researchers must use other theories and tools to explore generalized causal inferences of this type. (Shadish et al., 2002, p. 348, emphasis in original)

The most central problem is that using simple randomized control designs in field settings, such as schools, often introduces difficult issues related to the integrity of the intervention—the key element of internal validity. The CCSS tradition advocates designs such as cohort control, norm comparison, and time series precisely because they are much more likely to provide greater control over this element. Thus, these alternative designs could be seen as more internally valid than the randomized control group design within such settings.[13]

Comparing Conclusions of the Threshold and Inclusive Approaches

The mixed model analysis in Study 2 indicated that the WWC's threshold-related criteria and standards had little impact on estimates of the program's effectiveness. The estimates were unaffected by design characteristics, time of publication, presence and absence of stipulated confounds, and numerous other threshold-related characteristics. They were also similar across characteristics of the students and schools. Some might conclude that this suggests there should be few concerns regarding the use of these variables to limit the pool of evidence considered. In other words, if the criteria and standards have no impact on the estimates of effects, what would be the harm in using them?

There are, however, at least three reasons to question such a conclusion. First, as noted above, our analysis involved only one curriculum with highly consistent results across a wide range of settings. It is not clear that the same result would appear in examining interventions with more variable findings. In fact, given the very small number of studies that pass the thresholds the probability that a review would capture the extent of such variation is quite small. Second, because the threshold approach results in such a small pool of accepted studies, its conclusions would undoubtedly be based on a much less diverse sample than the inclusive approach. An inclusive approach would be much more likely to incorporate a range of settings, students, and communities that is more reflective of the larger population and thus heighten external validity of any summary judgment. Third, applying the limitations of the threshold approach could, in all likelihood, affect the accuracy of statistical results presented in a review. By definition, estimates of a given phenomenon, such as the effectiveness of an intervention, are more accurate (i.e., have smaller standard errors) when the sample used to develop that estimate is larger relative to the population. Thus, by limiting the number of studies included in a review, the margin of error is increased.

In developing summary reports of best evidence, the WWC uses aggregate data: averaging the effects within each study and counting the number of studies with positive, negative, or indeterminate effects to produce a categorical judgment. Our analysis highlighted several concerns with this method. First, because so few studies pass the defined thresholds, conclusions would most often be within the ''potential'' or ''mixed'' categories of ratings. Second, creating aggregate effects across a given study could potentially mask important variations across subgroups or other elements of a sample or analysis. Third, and perhaps the most important, is the nature of the rules themselves. To receive a positive (or negative) judgment, a set of studies regarding the intervention must have unanimous results—no contrary findings. Larger research literatures face particular challenges within the WWC rule-based system. As a simple result of statistical probabilities, interventions with a larger number of efficacy studies would be more likely to have at least one with contradictory

results. This would automatically preclude a positive (or negative) judgment on the literature, no matter what the full body of evidence indicated. Ironically, the presence of a larger body of evidence, seen as a hallmark of a mature science, works against the probability of acceptance within the review process.

Policy Implications

Future research should, of course, replicate our empirical analysis with other data sets and with the specific criteria and standards used by other review bodies. In the meantime, it would seem appropriate for groups that develop best evidence reviews, such as the WWC, to consider a number of policy changes to help ensure that summaries provide the most accurate information possible.

First, it would seem important to widen the net to include a broader sample of efficacy studies and indicators. At present, the WWC provides separate best evidence reports on literacy programs for students grouped by variables such as age, disability status, and language. Yet, users are often interested in the extent to which results replicate across multiple groups and settings. The WWC reports also focus on only certain outcome variables, primarily academic achievement. Yet, users may also be interested in other outcomes, such as student self-confidence and teachers' views of an intervention. Other aspects that could be central to accurate estimates of the impact of an intervention, such as measures of dosage, fidelity, and maintenance effects, are currently not part of the WWC's protocols. Yet, knowing the impact of these elements would be informative for potential users and, as explained above, provide more accurate estimates of effectiveness. Finally, rules that appear to eliminate large numbers of studies from consideration should be eliminated or sharply modified. Key among these would be the ''one-unit'' standard, which automatically disqualifies studies that include data from only one district or school, no matter how many schools or classrooms are involved or what types of controls are included. Another would be the requirement of minimal pretest differences on all variables, a standard that becomes very difficult to meet when the number of outcome variables is large or a sample is small.

Second, it would seem appropriate to consider methods of developing summary judgments other than a simple categorization of average effects. Using a more traditional summary analysis, such as the one utilized in Study 2, would counter the issues noted above and would also more closely parallel analytic strategies commonly used within the social sciences. Such analyses could provide estimates of the extent of variations in results by study- and design-related variables, including dosage and fidelity; factors related to external validity, such as those involving characteristics of the students, schools, and communities; and maintenance effects. This type of information is especially relevant for potential users, who would likely want to know how an intervention worked in situations similar to their own (Gargani & Donaldson, 2011; Granger & Maynard, 2015; Jacobs, Sisco, Hill, Malter, & Figueredo, 2012).

Finally, it would seem important to include the results of other summary analyses of a given intervention both within the analysis process and in the summary reports of best evidence. The efficacy literature includes a large number of meta-analyses and literature reviews of educational programs and other types of interventions. Such reviews can differ in their scope and content and systematic comparison of results across a variety of approaches can only increase users' confidence in the accuracy of a given summary (see, e.g., the works of Hattie, 2009 and Coughlin, 2014). Explicitly examining other reviews would also provide an important internal check for those developing the reviews to help ensure that their results are as accurate as possible.

Conclusion

It is important to emphasize the extent to which evaluators and social researchers, such as those who wrote to the Federal Register in 2003, support the intent of the EBP movement and the goal of

informing the public and policy makers of research findings. The various WWC criteria and standards were, no doubt, developed in good faith. Each, by itself, could potentially appear reasonable and appropriate. Yet, when taken together, they seem to have resulted in a system that drastically limits information provided to the public and the accuracy of the conclusions presented. They also could encourage researchers to restrict the questions they examine and approaches they use. Thus, in total, the concerns expressed by the majority of those responding to the Federal Register in 2003, appear to have been justified.

Authors' Note

Both authors are employed, on a part-time salaried basis, by the National Institute for Direct Instruction, which provides implementation support for the intervention examined in one of the studies reported in the article. The authors received no other financial support.

Acknowledgments

We thank Ashly Cupit for clerical assistance and Muriel Berkeley, Douglas Carnine, Christina Cox, Gary Davis, Siegfried Engelmann, Emily Hancock, Nicole Ngo, Robert O'Brien, Jerry Silbert, and Walter Wood for comments on earlier drafts of portions of the article. Any errors or conclusions in the article are the sole responsibility of the authors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material for this article is available online.

Notes

1. The What Works Clearinghouse (WWC) also may choose to describe results from studies that use a regression discontinuity design or single-case study, although this appears to be relatively uncommon. Given space limitations, our analysis only focuses on group designs.
2. We focus in this discussion on the standards most often used to eliminate studies with the threshold approach. Our discussion of the standards is limited due to space limitations and our major interest in contrasting the elements of the threshold approach to the Campbell, Cook, Shadish, and Stanley tradition. Other standards and criteria that have received criticism include those regarding areas such as attrition and acceptable measures of the dependent variables. An extended theoretical and statistical analysis of issues raised in this section is available upon request from the authors.
3. Other requirements to establish baseline equivalence include demonstrating equivalence separately for each outcome and the presence of either preintervention measures that are ''analogous'' to postintervention measures or the use of control variables specified by the WWC (2014, pp. 15–16).
4. We also made no attempt to check the accuracy of the listings or conclusions, and this could be an avenue for future work. Questions have been raised regarding the accuracy of WWC reports, including the studies listed and conclusions reached. One analysis of these concerns found examples involving 28 of the 93 interventions used in Table 1 (Wood, 2014; see also Confrey, 2006; Greene, 2010; Hempenstall, 2014; Shoenfeld, 2006; Sloane, 2008).

5. While there were no significant differences in the number or proportion of studies found eligible for review by topic area of the protocol, outcome of the review, or year of publication of the review, reviews for adolescent literacy identified significantly more efficacy studies to consider. Additional details regarding these variations are available on request from the authors.
6. The alternative titles included *DISTAR*, the name given to the program when it was first developed; *FUNNIX*, a computerized version of *Reading Mastery* (*RM*); *Horizons*, a very slightly modified version of *RM* designed for students with more advanced early literacy skills; and *Teach Your Child to Read*, a version designed for parents to use with their children. Studies of *Corrective Reading*, an accelerated version of *RM* designed for older students, were omitted unless the results of such studies had been combined with those of *RM*. The multivariate analysis included controls for these different titles and combinations of programs.
7. The WWC uses Hedge's g rather than Cohen's *d*. Both values could be computed for 31% of our comparisons. As would be expected given the very small differences in the formulas for the two measures, the average difference between the values was minimal (mean difference = .01).
8. The WWC uses only aggregated scores for their reports. We chose to use separate scale scores to allow examination of variability related to study characteristics.
9. Analysis of Research Questions 1 and 3 used studies as the unit of analysis, paralleling the decision-making logic of the WWC.
10. Of the 15 studies for which the pretest difference could not be calculated, three reported that there were no significant differences, seven adjusted for pretest scores in the calculations, three presented pretest data in earlier studies in the series of results, and only two gave no data or indication of the differences.
11. On average, the students in the intervention groups had substantially lower scores than those in the control groups, −.24 of a standard deviation (*SD*). However, there was substantial variation in these pretest differences, with a range of over 4 *SD* units (see Online Appendix B).
12. For example, the first edition of the WWC *Procedures and Standards Handbook* (2008) did not include the statistical requirements regarding the extent of pretest differences nor the "one-unit" rules requiring multiple schools and districts within an intervention group.
13. Interestingly, the WWC design criteria are more restrictive than Campbell and Stanley's definition of "true experimental designs." The latter includes the posttest-only control group design with random assignment within their list (1963, pp. 25–26), while the WWC criteria reject such designs. We have not been able to find a theoretical or empirical justification for the exclusion of these designs.

References

Adams, G., & Engelmann, S. (1996). *Research on Direct Instruction: 25 Years beyond DISTAR*. Seattle, WA: Educational Achievement Systems.

American Evaluation Association. (2003). *American Evaluation Association response to U.S. Department of Education, notice of proposed priority, federal registry RIN 1890-ZA00, November 4, 2003*. Retrieved from http://www.eval.org/p/cm/ld/fid=95

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science*, *1*, 31–49.

Biglan, A., Flay, B. R., Komro, K. A., Wagenaar, A. C., & Kjellstrand, J. (2012). *Adaptive time-series designs for evaluating complex multicomponent interventions in neighborhoods and communities*. Eugene, OR: Oregon Research Institute.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*, 125–230.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297–312.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Chen, H. T, Donaldson, S. I., & Mark, M. M. (2011). Validity frameworks for outcome evaluation. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing validity in outcome evaluation: Theory and practice. New Directions for Evaluation* (Vol. 130, pp. 5–16).

Confrey, J. (2006). Comparing and contrasting the National Research Council report on *Evaluating curricular effectiveness* with the What Works Clearinghouse approach. *Educational Evaluation and Policy Analysis*, *28*, 195–213.

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

Coughlin, C. (2014). Outcomes of Engelmann's Direct Instruction: Research syntheses. In J. Stockard (Ed.), *The science and success of Engelmann's Direct Instruction* (pp. 25–54). Eugene, OR: NIFDI Press.

Donaldson, S. I., & Christie, C. A. (2005). The 2004 Claremont debate: Lipsey vs. Scriven: Determining causality in program evaluation and applied research: Should experimental evidence be the gold standard? *Journal of MultiDisciplinary Evaluation*, *2*, 60–77.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (2009). *What counts as credible evidence in applied research and evaluation practice?* Los Angeles, CA: Sage.

Donaldson, S. I., Patton, M. Q., Fetterman, D. M., & Scriven, M. (2010). The 2009 Claremont debates: The promise and pitfalls of utilization-focused and empowerment evaluation. *Journal of MultiDisciplinary Evaluation*, *6*, 15–57.

Drake, R. E., Latimer, E. A., Leff, H. S., McHugo, G. J., & Burns, B. J. (2004). What is evidence? *Child and Adolescent Psychiatric Clinics of North American*, *13*, 717–728.

Fisher, R. A. (1925). *Statistical methods for research workers*. New York, NY: McGraw-Hill.

Fisher, R. A. (1935). *The design of experiments*. London, England: Oliver & Boyd.

Gargani, J., & Donaldson, S. I. (2011). What works for whom, where, why, for what, and when? Using evaluation evidence to take action in local contexts. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing validity in outcome evaluation: Theory and practice. New directions for evaluation* (Vol. 130, pp. 17–30). San Francisco, CA: Wiley.

Granger, R. C., & Maynard, R. (2015). Unlocking the potential of the "what works" approach to policymaking and practice: Improving impact evaluations. *American Journal of Education*, *36*, 558–569.

Greene, J. P. (2010). *What doesn't work clearinghouse*. Retrieved from http://educationnext.org/what-doesnt-work-clearinghouse/

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, England: Routledge.

Hempenstall, K. (2014). What works? Evidence-based practice in education is complex. *Australian Journal of Learning Difficulties*, *19*, 113–127.

Jacobs, W. J., Sisco, M., Hill, D., Malter, F., & Figueredo, A. J. (2012). Evaluating theory-based evaluation: Information, norms and adherence. *Evaluation and Program Planning*, *35*, 354–369.

Kazdin, A. E. (2004). Evidence-based treatments: Challenges and priorities for practice and research. *Child and Adolescent Psychiatric Clinics of North America*, *13*, 923–940.

National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. In R. J. Shavelson & L. Towne (Eds.), Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Oancea, A., & Pring, R. (2008). The importance of being thorough: On systematic accumulations of "What Works" in education research. *Journal of Philosophy of Education*, *47*, 15–39.

Office of the Federal Register, National Archives and Records Administration. (2005). Scientifically based evaluation methods, Department of Education, notice of final priority, RIN 1890-ZA00. *Federal Register*, *70*, 3586–3589.

Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York, NY: Basic Books.

Schmidt, W. P. (2014). The elusive effect of water and sanitation on the global burden of disease. *Tropical Medicine and International Health*, *19*, 522–527.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shoenfeld, A. H. (2006). What doesn't work: The challenges and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Research*, *35*, 13–21.

Sloane, F. (2008). Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher*, *37*, 41–46.

Slocum, T. A., Detrich, R., & Spencer, T. D. (2012). Evaluating the validity of systematic reviews to identify empirically supported treatments. *Education and Treatment of Children*, *35*, 201–233.

Slocum, T. A., Spencer, T. D., & Detrich, R. (2012). Best available evidence: Three complementary approaches. *Education and Treatment of Children*, *35*, 153–181.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*, *35*, 311–327.

Stockard, J. (2010). An analysis of the fidelity implementation policies of the What Works Clearinghouse. *Current Issues in Education*, *13*, 1–24. Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/398

Tallmadge, G. K. (1977). *The Joint Dissemination Review Panel ideabook*. Washington, DC: National Institute of Education and U.S. Office of Education.

Tallmadge, G. K. (1982). An empirical assessment of norm-referenced evaluation methodology. *Journal of Educational Measurement*, *19*, 97–112.

What Works Clearinghouse. (2008). *WWC procedures and standards handbook* (Version 1.0). Washington, DC: Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

What Works Clearinghouse. (2014). *WWC procedures and standards handbook* (Version 3.0). Washington, DC: Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Wood, T. W. (2014). *Examining the inaccuracies and mystifying policies and standards of the What Works Clearinghouse: Findings from a FOIA request* (NIFDI Technical Report 2014-5). Eugene, OR: National Institute for Direct Instruction.

Zvoch, K. (2012). How does fidelity of implementation matter? Using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation*, *33*, 547–565.