Examining the What Works Clearinghouse and Its Reviews of Direct Instruction Programs

Technical Report 2013-1

Jean Stockard, Director of Research
NATIONAL INSTITUTE FOR DIRECT INSTRUCTION
SPRING, 2013

Examining the What Works Clearinghouse and Its Reviews of Direct Instruction Programs: Executive Summary

The What Works Clearinghouse (WWC) is a federally funded program that evaluates educational interventions and provides summary ratings on its website. Scholarly literature reviews and meta-analyses are unanimous in concluding that a large research base indicates that Direct Instruction (DI) programs are highly effective. In contrast, the WWC has found very few studies of DI that meet its criteria for review and has concluded that there is little evidence to support the programs' efficacy.

This report provides an analysis of why the conclusions of the WWC regarding DI differ so markedly from the extant scholarly literature. The first section discusses issues related to criteria regarding exclusion or inclusion of studies and the WWC review procedures. Areas discussed include the ways in which reviews focus on narrow curricular programs, fail to examine or consider the characteristics of the programs, apply an arbitrary time limit to the included studies, and use standards for review that differ markedly from those generally used in the social sciences, excluding most field-based studies and those using advanced statistical methods. The second section examines the studies of DI programs that the WWC has found to meet their inclusion criteria, either with or without reservation. It documents serious errors in decisions regarding 4 of the 7 studies that were deemed as meeting their criteria "without reservation." The third and fourth sections analyze content of the WWC report on *Reading Mastery* and students with learning disabilities that was initially posted in July 2012, reviewing errors in inclusion and exclusion of studies. Over twenty research studies that could have been included in the WWC review are examined, detailing the design and conclusions of the studies, the effect size associated with their results, and reasons that the WWC might reject the study for inclusion.

The effect sizes are statistically analyzed, using mixed models and testing the hypothesis that the criteria used by the WWC provide a more accurate estimate of the efficacy of a curriculum. None of the study criteria used for inclusion/exclusion by the WWC was significantly related to the magnitude of effect sizes. In other words, there was no indication that the criteria used by the WWC to select or exclude studies from consideration were related to the reported results. The report concludes that the WWC procedures appear to result in a selective and inaccurate view of the DI literature, and this is probably the basis for the discrepancy between their conclusions and those of the research literature. It suggests that consumers would be well advised to consult sources of summary material other than the WWC and, especially, the well-conducted and highly regarded meta-analysis literature.

Table of Contents

Examining the What Works Clearinghouse and Its Reviews of Direct Instruction Programs[1]

The What Works Clearinghouse (WWC) is a federally funded program established in 2002 that evaluates educational interventions on the basis of the "rigor of research evidence" and provides summary ratings on its website. To date the Clearinghouse has produced seven reports on Direct Instruction (DI) curricular programs.[2] Scholarly literature reviews and meta-analyses are unanimous in concluding that a large research base indicates that DI programs are highly effective. In contrast, the WWC has found very few studies that meet its criteria for review and has concluded that there is little evidence to support the programs' efficacy. Given these sharply contrasting conclusions and the broad publicity given to the WWC's reports, it is important, from both a scholarly and a policy perspective, to understand more about why these discrepancies have occurred.

This report includes five sections. The first presents reflections on why the WWC procedures produce erroneous conclusions and why six of these procedures may have an especially severe impact on reviews of DI programs. The second section describes the 10 individual research studies of DI programs that the WWC concluded met their criteria for review. Serious errors in these decisions are described for a substantial proportion of these studies. The third section focuses on the WWC review of *Reading Mastery (RM)* and students with learning disabilities (LD), which was released in July 2012 and discusses the WWC's decisions regarding each of the 17 studies that were examined.[3] The fourth section analyzes 21 studies that could have potentially been included in the review of *RM* and students with

[1] The author thanks Douglas Carnine, Christina Cox, Sarah Haffner, Jerry Silbert, Piper Van Nortwick, Tina Wells, and Tim Wood for their assistance with elements of the preparation of this report. Any errors that remain are solely the responsibility of the author.

[2] The WWC issued reviews of DI programs with regard to ELL students in 2006 and early childhood education in 2007 (WWC, 2006, 2007a). Reports specifically targeted to *Corrective Reading* were issued in 2007 for beginning reading (WWC, 2007b) and 2010 for adolescent literacy (WWC, 2010a). Reports on *Reading Mastery* were issued in 2008 for beginning reading (WWC, 2008), 2010 for adolescent literacy (WWC, 2010b), and 2012 for students with learning disabilities (WWC, 2012).

[3] The WWC withdrew the original report on *RM* and LD students (WWC, 2012, published in July) in response to NIFDI's request for a quality review, but has indicated that it will reissue the report, eliminating one of the two articles (Herrera et al. 1997) to which NIFDI objected. (More details are below.) In mid-December 2012, the NIFDI office requested additional information on the decisions made in the reviews but has received no substantive response to this request.

LD, using preliminary results from a meta-analysis that the NIFDI Office of Research and Evaluation is preparing. A fifth section summarizes the analysis.

This report should be seen as an adjunct to other writings on the WWC and DI curriculum, including an extensive analysis of the 2008 WWC report on *Reading Mastery* (Stockard, 2008), the WWC's procedures with respect to implementation fidelity (Stockard, 2010) and acceptable research designs (Stockard, 2013), and reports regarding the most recent review of *RM* and students with LD (Stockard & Wood, 2012).

I. Six WWC Procedures and Their Impact on Reviews of DI Programs

There are a large number of empirical studies on the efficacy of Direct Instruction programs, but the What Works Clearinghouse has found very few that pass their criteria for inclusion. As noted above and described extensively in other documents (e.g., Stockard, 2008), the WWC's conclusions regarding the efficacy of the DI programs are in stark contrast to the various reviews and meta-analyses of the literature. These reviews have consistently found strong support for the programs' efficacy, often noting the large number of studies available and the large effects relative to other programs (e.g., Adams & Engelmann, 1995; Borman, Hewes, Overman, & Brown, 2003; Hattie, 2009; White, 1988). Six elements of the WWC procedures may have a special impact on reviews of DI materials.[4]

First, the WWC usually approaches its work by examining discrete programs rather than general curricula or approaches. The DI programs all are developed using the same methodology of careful analysis of the content to be taught and extensive attention to and testing of the structure, organization, wording, and sequencing of material.[5] Yet, the WWC treats each of the programs as a separate entity. In the most recent report, they used a comparison of two DI reading programs (*Reading Mastery* and *Horizons*), which found that both produced significantly greater growth than national norms, as a way to suggest that

[4] There are a number of additional problems with the quality of the WWC's procedures, such as the failure to revise and update reports on an on-going basis, their procedures of handling issues of implementation fidelity, and inconsistent quality and decisions across reviews. Because these problems probably affect all programs (although those with fidelity are more harmful to effective programs and aid less effective ones as shown in Stockard, 2010) and have been examined extensively by other authors, they are not discussed here. A thorough analysis of problems with WWC procedures should, of course, take these other issues into consideration.

[5] For instance, the reading programs—*DISTAR, Reading Mastery, Horizons, Funnix,* and *Corrective Reading*—are very similar in these technical elements. Their differences involve changes that were found to be appropriate for different audiences: *Horizons* and *Funnix* are developed for students who know letter names and shapes and thus these are introduced early in the sequencing. *Corrective Reading* is designed for older students and proceeds much more quickly to help students catch up with their peers. Reading *Mastery* is the current name for the *DISTAR* program.

one of them (and presumably the other if it is to be reviewed at a later time) actually had no effect. (See Stockard & Wood, 2012 for a more extensive discussion of this issue.) A much more appropriate way to review curricular programs, and one that is used by most researchers in the field, is to look at the broader curricular approach. The Direct Instruction programs are unique in employing such systematic and careful procedures in development and thus this part of the WWC approach is especially harmful to reviews of their curricula. An appropriate analysis of the research on Direct Instruction programs would include all of the works within an instructional area (e.g., for reading, *DISTAR, Reading Mastery, Horizons, Funnix,* and *Corrective Reading*). (Interestingly the WWC did use this broader approach in the ELL review (WWC, 2006, 2007a).)

Second, the WWC reviews do not appear to be based on an understanding of the theoretical and conceptual underpinnings of curricular approaches. As a result, they have made decisions regarding both inclusion and exclusion of studies that appear to be erroneous. Because the DI programs are much more firmly grounded in a theoretical and empirically substantiated body of work than other programs, this approach may especially impact the reviews of DI. As an example of faulty inclusion, and as described in more detail below, the reviews of *Corrective Reading* accepted a study that used elements of the program, even though the authors (Torgesen, Myers, et al., 2006; Torgesen, Schrim, et al., 2007) explicitly stated that their analysis should not be treated as an efficacy study of the curriculum. One of the most egregious examples of faulty exclusion is the decision to discard the RITE study, a multi-year study of the use of *Reading Mastery* in a large public school system (Carlson & Francis, 2002). The WWC chose to exclude this study from its review of *RM* for beginning reading, claiming that the inclusion of teacher training and behavioral management techniques, two key and essential elements of the DI approach, were additions to the curriculum and produced an inappropriate confound (see Stockard, 2008, pp. 10–12, for an extensive discussion of this decision).

Third, the WWC has set an arbitrary time limit of studies that it will consider, focusing only on works that have been published in the last 20 years and stating that this limitation helps ensure that findings are most relevant to today's students. This time limit appears to affect especially the reviews of DI materials, for the development of the programs and the smaller experimental studies that were used in this process occurred long before the cut-off date. Research articles continue to appear, but the later studies have tended to be examinations in real-life settings as well as studies on specific populations and variables related to improving implementations and applications. To date, the WWC has provided no empirical evidence that the way in which students learn has altered over the decades. The NIFDI research staff has found no other area in which findings from earlier eras are discounted or ignored if there is no empirical evidence of change in the underlying phenomena. An extensive discussion of this point is included in an analysis of the WWC's review of *Reading Mastery* (Stockard, 2008).

Fourth, the WWC has adopted "standards" regarding acceptable research designs, reserving their highest ratings to studies that employ random assignment. As explained elsewhere (Stockard, 2013), this approach contrasts sharply with the classic methodological literature and contemporary practices in the social sciences. Campbell and Stanley (1963) and their successors, authors of the most influential and widely cited books on research design, discuss the problems of employing random assignment in institutional settings such as schools. More importantly, they describe alternatives to random assignment that can counter these problems, alternatives that are both internally and externally valid. They also stress the importance of using a variety of research designs in the most recent edition of their work,

> 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. (Shadish, Cook, & Campbell, 2002, p. 30, emphasis in original)

In an extended discussion, Shadish, Cook, and Campbell (2002) discuss the importance of cumulating findings about a phenomenon using a variety of designs and a broad range of samples and measures. This perspective is, of course, parallel to the classic notion of cumulative knowledge within the Popperian tradition common to the scientific community.

Fifth, in addition to the preference for randomized studies the WWC requires that studies include pretest measures within given ranges and provide extensive information on attrition of subjects. This appears to effectively exclude many, if not most, studies that employ statistical analysis models, a standard way that contemporary social scientists use to adjust for variations among subjects in field settings. (Such reports often include these scores as control variables but don't provide the details the WWC seems to require.) The WWC standards also require a comparison group, apparently excluding norm comparison designs, even though such a design was traditionally promoted by the U.S. Department of Education (Tallmadge, 1977). Reviews are usually narrowly focused on an age or grade range and topic. They also may include restrictions to studies conducted in the United States and, as noted above, a limited time period. It is perhaps not surprising then that the WWC accepts only a small proportion of studies reviewed. (See the next section for a more complete analysis of the studies accepted for WWC review.) To the extent that studies of Direct Instruction materials are more likely than others to occur in field settings and thus involve

statistical analyses and approaches other than those used in small-scale laboratory settings, they may be especially affected by these restrictions.[6]

Finally, a sixth issue is the limitation of the most recent review of *Reading Mastery* to studies regarding students with learning disabilities, which artificially restricts the range of studies to be reviewed. The term "learning disabled" has not been consistently used during the last half century or even currently, and the criteria used to identify a student as learning disabled varies from one location to another, depending upon court cases and state laws. In other words, both the conceptual and operational definition of learning disability has altered over space and time. Yet, the WWC chose to restrict its review to studies that explicitly noted that more than half of the students in the samples had this "diagnosis." One could suggest that it would be reasonable to use a more inclusive definition, such as "struggling readers" or focusing on those exhibiting below-average achievement. These are, in fact, the students with whom we should be most concerned in developing remedial programs.

To summarize, several aspects of the WWC's review procedures may contribute to their erroneous reviews of the DI curriculum: 1) their focus on named curricular programs rather than broader curricular approaches, ignoring identical features of the most central elements of the DI reading programs; 2) their failure to understand or consider the characteristics embedded within curricular programs, such as the specific inclusion of teacher training and behavior management in the DI programs; 3) the use of an arbitrary time limit of studies to review, omitting at least half of the available corpus of material; 4) "standards" regarding acceptable study design that contrast markedly with the classic and contemporary methodological literature; and 5) "standards" for study characteristics that are often difficult for work conducted in field settings to meet. A sixth issue, specific to the review of *RM* and students with LD involves ignorance of the way in which the definition of LD has not been stable over space and time, resulting in the potential omission of large numbers of studies from review. These features have contributed to errors in the WWC's reviews of programs and, especially, in their decisions to exclude large proportions of studies from their reviews. The next section focuses on another very problematic area—the quality of decisions regarding the studies that are included in their reviews.

[6] Stockard (2013) discusses the way in which the research associated with DI is a prime example of the "phased model" of research, moving from rather small and highly focused controlled experimental designs in the early years of development to tests with more varied settings, subjects, and outcomes in later years. This progression is typical of a systematic and mature line of research, yet, ironically, these characteristics have hampered the ratings that DI receives from the WWC.

II. Ten Studies Accepted for Review by the WWC

The WWC reports that it has reviewed over 200 studies of DI programs in developing its reviews. Yet they have found only seven that fully met their criteria for review and an additional three that met their criteria "with reservation" (Stockard, 2012). As described more fully below, NIFDI's Office of Research and Evaluation is conducting a meta-analysis of studies of the Direct Instruction curriculum and has identified several hundred more studies for review.[7] Because the WWC's conclusions are based on such a small proportion of the extant literature, it is important to examine the characteristics of the selected studies. The sections below give citations to and a brief description of each of the studies that the WWC has accepted and an assessment of whether or not it was appropriate to accept the study for review. The analysis suggests that there were problems in a number of the decisions. The studies that were "accepted without reservation" (fully met WWC criteria) are examined first, followed by those that were "accepted with reservation."

A. Accepted Without Reservation (7 studies)

The WWC accepted 7 of the 200-plus studies that were reviewed "without reservation." The analysis suggests that decisions regarding 3 of the 7 studies were correct, but that 4 of these decisions were not appropriate.

A.1 Decisions that were appropriate (3 studies)

1) Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *The Journal of Special Education, 34*(2), 90–103.

2) Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *The Journal of Special Education, 36*(2), 69–79.

These two studies were accepted in the review of the programs for English language learners. They were rejected, however, for the reviews of *RM* for beginning reading and for

[7] The WWC reports that they looked at a total of 228 studies of DI programs. The 7 that "fully" met the inclusion criteria represent 3.1% of this total; when the 3 that met the criteria "with reservation" are included, the percentage rises to 4.4%. To date the NIFDI research office has identified 557 studies for review for the meta-analysis. While a large number of the studies may not meet criteria for inclusion in the final analysis, the fact that the NIFDI office has identified over twice as many studies as the WWC suggests that their search of the literature may have been substantially less thorough than would be appropriate. Each of the other meta-analyses of the DI literature has also identified many more studies for review. Hattie's (2009) meta-analysis of meta-analyses summarized the results of four meta-analyses that included DI, incorporating 304 studies and 597 separate effects.

adolescent literacy. The latter decision is understandable given the age of students in the study. The NIFDI Research Office has, however, objected to their rejection for the former and continues to believe that this was not appropriate (Stockard, 2008). As detailed in a later section, this study would also have been appropriate to include in the study of *RM* with students with LD. Gunn and associates have published a third article on the data (Gunn et al., 2005), providing additional follow-up. This was not used in the ELL report, perhaps because of the date of publication, but was rejected for inclusion in the *RM* reports for beginning reading and adolescent reading as well as for the two reports on *Corrective Reading*. All three of the studies will be included in the meta-analysis. (In Table 2 these are the Gunn studies 6, 7, and 8.)

3) Stockard, J. (2010*). Fourth graders' growth in reading fluency: A pretest-posttest randomized control study comparing* Reading Mastery *and* Scott Foresman Basal Reading Program. Eugene, OR: National Institute for Direct Instruction.

This small study was included in the report on *RM* for adolescent literacy. It involved random assignment of high achieving students to receive *RM* or to continue in the school's usual curriculum. Its inclusion seems appropriate, and it will be included in the meta-analysis.[8] (This is study 20 in Table 2.)

A.2 Accepted without reservation, but the decision was probably in error (4 studies)

It could be suggested that the other 4 studies that were accepted without reservation should not have been included, at least in the way in which the WWC chose to interpret them.

4) Cooke, N. L., Gibbs, S. L., Campbell, M. L., & Shalvis, S. L. (2004). A comparison of *Reading Mastery Fast Cycle* and *Horizons Fast Track A–B* on the reading achievement of students with mild disabilities. *Journal of Direct Instruction, 4*(2), 139–151.

This study compared the efficacy of two DI reading programs, *Reading Mastery* and *Horizons*. It was accepted without reservation for the WWC review of *RM* for students with learning disabilities, probably because it included random assignment of students to treatment. The study compared students' growth over time to national and state norms and between the two programs. Students in both programs had significantly stronger growth than others in the state and nation and growth rates that were similar to each other. The WWC concluded that this result indicates that *RM* is not effective. However, it would be more appropriate to use the study as a norm comparison design, comparing *RM* students with

[8] A later analysis of the data in this study focused on growth in reading comprehension. Although this was submitted to the WWC shortly after completion it has not, to date, been added to their review.

national and state norms. As explained more fully in Stockard and Wood (2012), the logical and appropriate conclusion should be that *RM* (and *Horizons*) produced stronger growth than national norms and that this indicates effectiveness of both programs. The study will be included in the meta-analysis, but as two different norm comparison designs. (This is study 3 in Table 2.)

5) Herrera, J. A., Logan, C. H., Cooker, P. G., Morris, D. P., & Lyman, D. E. (1997). Phonological awareness and phonetic-graphic conversion: A study of the effects of two intervention paradigms with learning disabled children. Learning disability or learning difference? *Reading Improvement, 34*(2), 71–89.

This study was originally accepted for the review of *RM* for students with learning disabilities (WWC, 2012). In this study all students received *Reading Mastery* as part of their usual classroom experience. Half of the students received an additional period of phonics-related instruction each day from their classroom teachers. The other half did not receive additional instruction.

Teachers were randomly assigned to treatment condition. As would be expected, students with extra instruction had stronger growth over time. The WWC originally concluded that this indicated that *RM* was ineffective, but has now apparently reversed this decision. (It has been included in this listing because it illustrates the problems with their review procedures.) It will not be included in the meta-analysis because there was no group in the analysis that did not have *RM*.

6) Torgesen, J., Myers, D. Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., et al. (2006). *National assessment of title I. Interim report. Volume II. Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers.* Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

7) Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., et al. (2007). *National assessment of Title I. Final report. Volume II. Closing the reading gap: Findings from a randomized trial of four reading interventions for striving readers (NCEE 2008-4013).* Washington, DC: National Center for Education Evaluation and Regional Assistance.

These reports were accepted for the WWC reviews of *Corrective Reading (CR)*. The "interim" report was accepted for the review of *CR* for beginning readers (WWC, 2007b) and both the final and interim reports were accepted for the review of *CR* for adolescent literacy (WWC, 2010a). The study was nicely designed with random assignment of schools and students to treatment. However, the study is not a test of the efficacy of the *CR* program and should not

have been included. On p. ix of the 2006 *Interim Report*, Torgesen and associates noted that

> *Corrective Reading* and *Wilson Reading* [another program used in the study] were modified to fit within the first of these classifications [used in the study]. The decision to modify these two intact programs was justified both because it created two treatment classes that were aligned with the different types of reading deficits observed in struggling readers and because it gave us sufficient statistical power to contrast the relative effectiveness of the two classes. Because *Corrective Reading* and *Wilson Reading* were modified, results from this study do not provide complete evaluations of these interventions; instead, the results suggest how interventions using primarily the word-level components of these programs will affect reading achievement.

In other words, the authors explicitly stated that the study should not be used to evaluate the program. The WWC report notes that the program was not implemented as designed, but chose to ignore the stipulations of the authors. The studies will not be included in NIFDI's meta-analysis because the programs involved modifications of *Corrective Reading* and the authors explicitly noted that they should not be used in this way.[9]

B. Accepted With Reservations (3 studies)

The WWC accepted 3 of the 200-plus studies that were reviewed "with reservations."

8) Cole, K. N., Dale, P. S., & Mills, P. E. (1991). Individual differences in language delayed children's responses to direct and interactive preschool instruction. *Topics in Early Childhood Special Education, 11*(1), 99–124.

9) Cole, K. N., Dale, P. S., Mills, P. E., & Jenkins, J. R. (1993). Interaction between early intervention curricula and student characteristics. *Exceptional Children, 60*, 17–28.

These two studies were accepted "with reservations" by the review of DI programs for early childhood education. (Both studies report on the same sample.) They were accepted "with reservations" rather than "without reservation" because of "severe attrition" to the sample over time. It was appropriate to include these studies and they will be in the meta-analysis.

[9] This decision is especially ironic given the WWC's failure to include the analysis of the RITE study (Carlson & Francis, 2002) in their analysis of *RM* and beginning reading, citing, as described above and in Stockard (2008) the supposed modification of the program through the inclusion of teacher training and behavioral reinforcement.

10) Yu, L., & Rachor, R. (2000, April 24-28). "The two-year evaluation of the three-year Direct Instruction program in an urban public school system," Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ED 441831)

This study was accepted, with reservations, for the WWC review of *RM* for adolescent literacy (WWC, 2010b). Schools and students in this study were matched on a variety of risk measures, and those exposed to DI were compared to those with exposure to another, unnamed curriculum. It did not receive the highest rating because there was no random assignment. The paper will be included in the forthcoming meta-analysis of all studies of Direct Instruction.[10]

To summarize, since beginning its work more than a decade ago, the WWC has concluded that only a small fraction of the studies that it has examined either fully or partially ("with reservations") met its criteria for review. Only 10 studies of DI programs, out of hundreds that are available, were accepted. However, careful examination of these 10 studies indicates serious errors in 4 of the 10 decisions. The next section involves a more in-depth look at the studies that the WWC reviewed in its July 2012, analysis of *Reading Mastery* and students with learning disabilities.

III. The WWC Study of *Reading Mastery* and Students with Learning Disabilities

This section, and the next, focus on issues related to the most recent (2012) WWC review, which examined the use of *Reading Mastery* with students with learning disabilities. This section examines the studies that the WWC listed in its report, while the next (section IV) looks at those that should potentially have been included. The original WWC review reported that 17 studies had been examined.[11] These studies are listed below, categorized by the decision reached by the WWC and a brief summary of an analysis of the articles. The discussion is divided into those that were considered by the WWC and accepted (A), and those that were rejected (B), further separating the analysis in this second section by the

[10] The NIFDI Research Office has not been successful in finding a published version of this paper or a report of the third-year evaluation of the program even though the title implies that such an evaluation should be available.

[11] The July 2012 report regarding *RM* and students with learning disabilities, which has been withdrawn from the WWC's website, also included the Herrera et al. study described above (section II, study 5) as fully meeting criteria. Communications received from the WWC indicate that a revised report will continue to use all of the studies in the original report but will move the Herrera et al. piece to a category indicating that it did not meet criteria. Thus, this study is listed in a category of works that do not meet criteria because the intervention is unclear (this section, B.3, study 11).

reasons listed for rejection. As explained in other documents (Stockard & Wood, 2012; Stockard 2012), there appear to have been numerous problems with this most recent, 2012, WWC report.

A. Studies That Were Considered by the WWC and Accepted (1 study)

1) Cooke, N. L., Gibbs, S. L., Campbell, M. L., & Shalvis, S. L. (2004). A comparison of *Reading Mastery Fast Cycle* and *Horizons Fast Track A–B* on the reading achievement of students with mild disabilities. *Journal of Direct Instruction, 4*(2), 139–151.

This study compared results with *RM* and *Horizons*. *Horizons* is a modified version of *RM*. The article describes 17 common features of the development and construction of the programs and only two differences: In contrast to *RM, Horizons* uses letter names to prompt letter sounds and uses capital letters in the first reading lessons. As would be expected, given these extensive similarities, Cooke and associates found that both groups in their study made gains over time that were significantly greater than national norms. The WWC used these findings to suggest that *RM* was not effective and has, apparently, decided to retain this conclusion. However, as noted above, this conclusion does not appear to be appropriate. The article is included in the meta-analysis results summarized below with what is a more appropriate analysis and interpretation (two norm comparison designs—one for *Horizons* and one for *RM*). (This is study 3 in Table 2.)

B. Studies That Were Considered by the WWC and Rejected for Inclusion (16 studies, 4 reasons)

Sixteen studies were considered by the WWC and rejected for inclusion. Four general reasons were given for these decisions: 1) less than half of the students in the study were identified as being learning disabled (n = 4), 2) a single subject or similar design was used (n = 4), 3) results could not be attributed solely to the intervention because of some confounding element (n = 2), and 4) the study was not an efficacy test (n = 6). Each of these studies is described below.

B.1 Rejected because less than 50% of the students had LD (n = 4 studies)

2) Butler, M. T. (2001). *Comparison of the effects of direct instruction and basal instruction on the reading achievement of first-grade students identified as students with reading difficulties* (Doctoral dissertation, University of Alabama, 2001). Dissertation Abstracts International, 62 (09A), 203-3002.

Although the NIFDI Research Office was unable to get the full copy of the dissertation before completing the analysis reported here, the abstract provides no reason to discount the interpretation of the WWC.

3) Kanfush, P. M., III (2010). *Use of direct instruction to teach reading to students with significant cognitive impairments: Student outcomes and teacher perceptions*. (EdD Dissertation, West Virginia University) Ann Arbor, MI: ProQuest, LLC.ERIC-ED 521267.

The NIFDI Research Office was unable to obtain a full copy of this dissertation before completing the analysis reported here. However, from the abstract it appears that there was no non-DI group, so no comparison would have been possible.

4) O'Connor, R. E., Jenkins, J. R., Cole, K. N., & Mills, P. (1993). Two approaches to reading instruction with children with disabilities: Does program design make a difference? *Exceptional Children, 59*(4), 312-323. (The WWC also reported looking at a 1992 unpublished manuscript with the same title.)

The study involved a transitional kindergarten program for children with disabilities. The term "learning disabled" isn't included in the list. Language, cognitive, motor, and socio-developmental delays are noted. This study will be included in the meta-analysis. (This is study 13 in Table 2.)

5) SRA/McGraw-Hill. (2009). *A report on the effects of SRA/McGraw-Hill's* Reading Mastery, Signature Edition: *A response to intervention solution*. Desoto, TX: Author.

As noted below and in the next section, there are a number of case studies that SRA/McGraw-Hill has published. Analyses of those that could have potentially been included in this review are in the discussion below and these reports will be included in the meta-analysis. (In Table 2, studies 15–18 are SRA studies. This 2009 report is not among them.)

B.2 Single subject or one unit per condition; No comparison group
(n = 4 studies)

6) Fitzpatrick, E., McLaughlin, T. F., & Weber, P. (2004). The effects of a first day and second day reads on reading accuracy with *Reading Mastery III Textbook B* for a fifth grade student with learning disabilities. *International Journal of Special Education, 19*(1), 56–63.

This uses a single subject design. It is appropriate to omit the study.

7) Lovett, M. W., Steinbach, K. A., & Frijters, J. C. (2000). Remediating the core deficits of developmental reading disability: A double-deficit perspective. *Journal of Learning Disabilities, 33*(4), 334.

It is unclear why this study was included in the listing. It appears to have used elements of *Reading Mastery* and *Corrective Reading* in one of the treatment groups (see p. 337) but incorporated them into another instructional system. Thus, it is not a test of the efficacy of either *RM* or *CR*. The WWC's classification is rather odd because, in fact, there were comparison groups. The study will not be included in the meta-analysis because it does not involve a DI program.

8) SRA/McGraw-Hill. (2006a). *Exceptional education and regular education students excel with Direct Instruction*. Retrieved from SRA website: http://www.mheresearch.com/assets/products/c9f0f895fb98ab91/iredell_statesville_schools.pdf

9) SRA/McGraw-Hill. (2006b). Reading Mastery, Corrective Reading *help students with disabilities achieve significant academic growth*. Retrieved from SRA website: https://www.sraonline.com/download/DI/EfficacyReports/Clover_DI.pdf.

The WWC rejected these two SRA reports saying there was no comparison group. However, as shown below, the reports fit the classic definition of a cohort comparison design and should have been included. (It is possible, however, as noted with the SRA report described earlier in this section under B.1, that the WWC would have also rejected the report because it wasn't clear that more than 50% of the subjects had LD. Students in the second report also had *Corrective Reading*.) (The 2006a SRA study is study 15 in Table 2.)

B.3 Result cannot be attributed solely to the intervention (n = 2 studies)

10) Earheart, L.S. (2002). *The efficacy of the SRA reading program for disabled learners as measured by the Terra Nova achievement test* (Doctoral dissertation, Tennessee State University, 2002). Dissertation Abstracts International, 63 (08A), 57-2823.

Although the NIFDI Research Office was unable to get the full copy of the dissertation before completing the analysis reported here, the abstract provides no reason to discount the interpretation of the WWC.

11) Herrera, J. A., Logan, C. H., Cooker, P. G., Morris, D. P., & Lyman, D. E. (1997). Phonological awareness and phonetic-graphic conversion: A study of the effects of two intervention paradigms with learning disabled children. Learning disability or learning difference? *Reading Improvement, 34*(2), 71–89.

This was originally accepted for the WWC analysis, but upon NIFDI's request, they have apparently removed it from the list of accepted studies. It compares *RM* with *RM* plus another program. The students in the two groups did not have equivalent instructional times, and all students had *RM*. Although the NIFDI office does not know what final decision the WWC has reached, placement in this category would be appropriate.

B.4 Not an efficacy test (no comparison to non-DI procedures) (6 studies)

One of the works in this group examines slight variations in the implementation (Frankhauser et al., 2001), while the others are reviews of the literature. The decisions regarding these works appear appropriate.

12) Frankhauser, M. A., Tso, M. E., & Martella, R. C. (2001). A comparison of curriculum-specified reading checkout timings and daily 1-minute timings on student performance in *Reading Mastery*. *Journal of Direct Instruction, 1*(2), 85–96.

This single-subject design examined ways in which various additions to standard DI procedures altered results.

The following five works are literature reviews of various types, and it is not clear why they were included in the original list:

13) Kinder, D. Kubina, R., & Marchand-Martella, N. (2005). Special education and direct instruction: An effective combination. *Journal of Direct Instruction, 5*(1), 1–36.

14) Scammacca, N., Vaughn, S., Roberts, G, Wanzek, J., & Torgesen, J. K. (2007). *Extensive reading interventions in grades K–3. From research to practice*. Portsmouth, NH: Center on Instruction.

15) Schieffer, C., Marchand-Martella, N. E., Martella, R. C., Simonsen, F. L., & Waldron-Soler, K. M. (2002). An analysis of the *Reading Mastery* program: Effective components and research review. *Journal of Direct Instruction, 2*(2), 87–119.

16) Stewart, R. M., Benner, G. J., Martella, R. C., & Marchand-Martella, N. E. (2007). Three-tier models of reading and behavior: A research review. *Journal of Positive Behavior Interventions, 9*(4), 239–253.

17) Swanson, H. L. (2011). Learning disabilities: Assessment, identification and treatment. In M. A. Bray & T. J. Kehle (Eds.), *The Oxford Handbook of School Psychology* (pp. 334–350). New York: Oxford University Press.

C. Summary

Of the 17 studies examined by the WWC for the review of *RM* for students with learning disabilities, classifications of four should be questioned: the decision to retain Cooke et al. (2004) as a comparison of *RM* and *Horizons* and the rejection of two SRA case studies (2006a, 2006b) and the O'Connor, et al. (1993) study. The next section of this report describes the studies that have been found for the meta-analysis that appear to be pertinent to the topic of the WWC's review of *RM* and students with learning disabilities.

IV. An Analysis of Studies That Could Have Been Included in the Review of *RM* and Its Use with Students with Learning Disabilities

This section examines studies of *Reading Mastery* that could have been included in the review of *RM* and students with LD. The first section describes the studies, and the second includes a brief statistical analysis, looking at the way in which the various WWC criteria are related to variations in effect sizes. If the criteria are important and valid to include, they should be related to variations in effect sizes, and this admittedly brief analysis addresses that issue.

A. Studies That Could Have Been Included (21 studies)

The studies listed below come from NIFDI's preliminary work on the meta-analysis of Direct Instruction efficacy studies. The list was limited to those that include *Reading Mastery*, although, as noted above, a more complete and accurate analysis would also include works that used *Corrective Reading* and other DI reading programs. Unlike the WWC analysis, articles were included that did not use the term "learning disabled," but had samples that focused on "struggling readers" or students scoring below normative levels. The many studies that focused on entire schools that had students with low levels of achievement, which also could be relevant, were not included. (The vast majority of students in these settings are "struggling readers" and scoring below their ability.) Most important, because the meta-analysis work is on-going, there are undoubtedly a number of other works that might be relevant. Those that are listed, however, should provide a good sampling of the literature base.[12]

[12] Four additional studies that might be applicable were identified, but at the time of the analysis the NIFDI Research Office had not yet been able to obtain copies: Herb, M. H. (2005). *The effects of* Reading Mastery *for students with learning disabilities.* Unpublished master's thesis. Pennsylvania State University, Philadelphia, PA; Lutz, A. R. (2004). *The effectiveness of the* Reading Mastery *reading program when teaching learning support students how to read.* Unpublished master's thesis, Gratz College, Melrose Park, PA; Ocokoljich, E. D. (1997). *The effects of* Reading Mastery I *and* II *on the reading achievement of first and second grade students identified as having low phonological awareness skills.* Unpublished master's thesis. University of Wisconsin–Madison, Madison, WI; and Sprinkman, A. (2001). *A comparison of reading achievement made by LD and low*

The studies are listed in alphabetical order (by author). The design of each study is described, including the associated effect size, calculated as the difference between the means divided by the common standard deviation (Cohen's d). Also included, for the works not in the WWC listing, is a brief speculation about why the WWC might have rejected the study if it had been included in their list. (Tables that summarize this information are included with the statistical analysis in this section. The numbers for the 21 studies described below correspond to the numbers in Table 2.)

1) Bowers, W. M. (1972). *An evaluation of a pilot program in reading for culturally disadvantaged first grade students*. (Doctoral Dissertation, University of Tulsa).

All of the participants scored below the 25th percentile of the Metropolitan Reading Readiness test. There were 8 classrooms in 4 schools, with 1 experimental and 1 control classroom in each school. Half of the classes in each group were self-contained and half were compartmentalized. *DISTAR* was used with the experimental group for 4.5 months, and the control group received the school's usual curriculum. Students were randomly assigned to treatment. Assessments were the Gates McGinitie Reading Test subtests of vocabulary and comprehension. Four effect sizes were calculated (two for compartmentalized and two for self-contained students). Results favored *DISTAR* for the self-contained students (average d = .52), but did not favor *DISTAR* students for the departmentalized students (d = –.22). The overall average d for the study was .15.

The WWC would probably reject this study because of the publication date and because it used *DISTAR* rather than *RM*. It also does not use the term "learning disabled."

2) Branwhite, A. B. (1983). Boosting reading skills by Direct Instruction. *British Journal of Educational Psychology, 53,* 291–298.

The students in this study might be classified as learning disabled in today's terminology for they had average ability, but low reading achievement. The students had a mean IQ of 91.64 (range 74 to 108), but were, on average, two and one-half years below grade level in reading skills. Seven of the subjects received *DISTAR Reading II,* and 7 received "diagnostic-prescriptive remediation" (DPR), described as a "phonic loading" program. Instruction occurred for 35 minutes each day, by the same teacher, for 110 days. After that time,

---

IQ students using a Direct Instruction reading program. Unpublished master's thesis. Cardinal Stritch University, Milwaukee, WI. In addition, the following study could potentially be used but the NIFDI office was unable to find enough information for the meta-analysis: Kuder, S. J. (1990). Effectiveness of the *DISTAR* reading program for children with learning disabilities. *Journal of Learning Disabilities, 23*(1), 69–71. There are undoubtedly a number of other studies that could be included and it is hoped that the final version of the meta-analysis will use such a complete list.

because of the significantly stronger growth of the *DISTAR* children, the DPR intervention was stopped and all students received *DISTAR*. The effect size was 1.71.

The WWC would probably reject this study because of the publication date, because it used *DISTAR* rather than *RM*, because it did not use the term "learning disabled," and/or because it occurred outside the U.S.

3) Cooke, N. L., Gibbs, S. L., Campbell, M. L., & Shalvis, S. L. (2004). A comparison of *Reading Mastery Fast Cycle* and *Horizons Fast Track A–B* on the reading achievement of students with mild disabilities. *Journal of Direct Instruction, 4*(2), 139–151.

As noted above (in section II as well as in section III), this article compared student gains over time in *Reading Mastery* and *Horizons*. Data came from 30 students, all "identified as having mild disabilities and … eligible for specialized reading instruction from a special education teacher in a resource room setting." The article explicitly notes that half of the students had a learning disability. Pre- and posttest measures at the beginning and end of the school year were the Woodcock-Johnson Revised and the North Carolina Literacy Assessment. As noted earlier, the WWC chose to focus on the lack of differences in change between the two programs and concluded that *RM* was not effective. Given the strong similarities of the two programs, a more appropriate approach would be to use a norm comparison design and ask if the students made gains over time that were greater than other students in the nation (or state with respect to the NCLA). Effect sizes for the *Reading Mastery* students ranged from .20 to .71, with an average of .34.[13]

The WWC has apparently chosen to reject this logic, preferring to treat *Horizons* and *RM* as separate programs. It is also not clear that they would accept a norm comparison design. However, the other aspects of the study seem to have been acceptable.

4) Francis, B. J. (1991) *Matching reading programs to students' needs: An examination of alternate programming using a direct instruction program in the regular classroom* (Master's thesis, Simon Fraser University).

The study took place in 6 schools in a suburb of Vancouver, B.C. Three schools implemented *RM* and 3 retained their usual program of instruction. Independent observers determined that the schools were similar in SES and staff effectiveness. Subjects were students in Grades 3 to 6 who had been nominated by their teachers as having problems with reading.

[13] The effect sizes reported by Cooke et al. (2004), which were calculated by dividing the difference between the pre and post means by the s.d. at pretest (p. 147), are used in the analysis below. An alternative method would use the s.d. of the national or state tests. These calculations consistently yielded slightly larger effect sizes, so the more conservative estimates derived by the authors' methods were used.

Pretests were administered in October and posttests in May on a variety of measures: Gates McGinitie Reading Test, Canadian Edition; Informal Reading Inventory, the Basic Reading Inventory (BRI), and Students' Perception of Ability Scale (SPAS). There were 162 students in the experimental group and 99 in the control group for a total of 261 students. Twenty effect sizes were calculated (5 measures and 4 grade levels). They ranged from –.70 to .89, with a mean of .11. While pretest differences were generally small (average = –.17, in favor of the control group), some were larger and this variation is considered in the statistical analysis of results reported below.

WWC might reject this study because of the year of publication, because it doesn't mention learning disability, and/or because it did not take place in the U.S. They might also reject it because some of the pretest differences exceeded .5 of a standard deviation.

5) Gersten, R. M., & Maggs, A. (1982). Teaching the general case to moderately retarded children: Evaluation of a five year project. *Analysis and Intervention in Developmental Disabilities, 2,* 329–343.

This study looks at the impact of instruction in *DISTAR* on students scoring in the "high-moderate range of retardation" (p. 332). It uses a norm-referenced comparison design, noting that this is "an evaluation research design advocated by Tallmadge (1977) for evaluations conducted for the U.S. Department of Education" (p. 332). The sample includes 12 subjects who were in the program for 5 years. Measures included the Stanford-Binet Intelligence Test, the Baldie Language Ability test, and the Neale Analysis of Reading. The study took place in Australia. The logic of the norm-referenced comparison design, comparing the change over time to the normative population, was used to calculate effect sizes, with posttest mean scores corrected for regression to the mean. The calculated effect size using the population s.d. (16) was .36, while the effect size using the sample s.d. (5.4) was 1.08.

The WWC may reject this article because of the year of publication, because it doesn't mention learning disability, because it used *DISTAR* rather than *RM*, because it did not take place in the U.S., and/or because it used a norm comparison design.

6) Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *The Journal of Special Education, 34*(2), 90–103.

7) Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *The Journal of Special Education, 36*(2), 69–79.

8) Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education, 39*(2), 66–85.

These three articles reported on the same study, using different follow-up time periods. Children in Grades 1–3 in 9 elementary schools in 3 districts were screened on reading (or pre-reading) skills. DIBELS measures were used for screening and for assessment as well as the Woodcock-Johnson Tests of Achievement (letter word ID and word attack subtests). Students were grouped by ethnicity and grade and rank ordered by pre-reading scores. Then participants were matched and randomly assigned to treatment or control. Supplemental instruction using *RM* (for Grades 1–2) or *CR* (for students in Grades 3 or 4) was provided for the experimental group. The 2000 article reports results at the end of the first and second year. The 2002 study reports results one year after the end of the intervention, and the 2005 article reports results 2 years after the end of the study. In all cases the *RM* and *CR* students had significantly stronger gains than the control group. For the 2000 report, there were 24 calculated effect sizes, ranging from 0 to .79, with an average of .34; for the 2002 report, there were 15 effect sizes, ranging from .05 to .74, with an average of .31; and for the 2005 report there were 10 effect sizes, ranging from .03 to .40, with an average of .23. Some of the calculated effects were based on data aggregated from other reports and this is addressed in the statistical analysis reported below.

This group of studies is very strong methodologically and it is not clear why it was not included in the list of works that were reviewed. The WWC may reject it for inclusion because it does not specifically mention "learning disabilities," instead focusing only on students who are below grade level. They may also reject it because part of the intervention involved behavioral training, although the authors are firm in noting that this should not be an issue. And they may reject it because students received either *RM* or *CR* depending upon their grade level or because pretest data weren't included in each of the reports. The last criticism should not be an issue because students were randomly assigned. When data on pretest scores were given (for some comparisons in 2005) there were no differences, with all effect sizes being smaller than .10 in absolute value.[14]

9) Kamps, D., Wills, H., Greenwood, C., Thorne, S., Lazo, J., et al. (2003). Curriculum influences on growth in early reading fluency for students with academic and behavioral risks. *Journal of Emotional and Behavioral Disorders, 11*(4), 211–224; Reprinted in 2004 in *Journal of Direct Instruction, 4*(2), 189–210.

[14] The 2000 and 2002 reports were included in the WWC's report regarding *RM* for ELL students (see section I). However, the studies were rejected for inclusion in the reviews for beginning reading and adolescent literacy for both *CR* and *RM* because they didn't fit the protocol (adolescent literary) or because there was a second intervention (the behavioral intervention) in the design (beginning reading). Only the 2002 and 2005 studies are mentioned in the beginning reading review.

This study looked at growth in reading skills over time of students in kindergarten through second grade who had been identified for behavioral and academic risk. Students in 5 schools participated in the study. One school used *Reading Mastery*, one used *Success for All*, and the other 3 used literature-based programs (n = 111 students for *RM*, 107 for SFA, and 164 for literature-based). Growth in reading skills, as measured by indicators in the DIBELS system, was examined over a three-year period, using standard multivariate (growth model) techniques. Results indicated that students in *RM* had stronger gains in reading fluency than those in the other programs. Effect sizes were calculated using the end-point performance for each curriculum group and standard deviations obtained from Good, Wallin, Simmons, Kame'enui, and Kaminski (2002). Six effect sizes were calculated, and they ranged from .19 to .73 with an average of .39.

This study is quite strong methodologically, and it is unclear why the WWC would reject it. These are possible reasons they might use: The article does not explicitly mention "learning disabled" although all of the students were defined as at risk. There are no comparisons of pretest scores, although all students had the same screening procedures and the multivariate analyses adjust for any differences. The students were also selected for behavioral issues and they might consider that element to be a confounding factor, even though that was not the focus of treatment. Finally, they may object to having both *CR* and *RM* as interventions (although this varied by the grade of the student). Note that this article was not included in the list of studies reviewed. (It was rejected for the WWC's review of *RM* for adolescent literacy in 2009 because it was out of the age range for the review, but it was not included in the list of studies in the 2008 review of *RM* for beginning reading.)

10) Marston, D., Deno, S. L., Kim, D., Diment, K., & Rogers, D. (1995). Comparison of reading intervention approaches for students with mild disabilities. *Exceptional Children, 62*(1), 20–37.

In this study 37 special education resource room teachers were randomly assigned to 1 of 6 different teaching approaches. There were 176 students in the study, described as having "mild disabilities." Teaching approaches were DI (described as the Engelmann Becker, SRA method), application of DI methods to a Holt basal reader, reciprocal teaching, peer tutoring, computer-assisted instruction, and effective teaching. Teachers received training in the method to which they were assigned, and an additional control classroom was also assessed. District curriculum-based measurement (CBM) reading probes were transformed to standard scores using district norms. Pretest score was median number of words read correctly on three passages in week 1 and posttest was median score in week 10. The DI group did not do significantly better in this analysis. The six calculated effect sizes ranged from –.84 (in favor of the control group) to .01, with an average of –.48. In the two comparisons with the largest absolute value of effect size, the pretest scores differed by

more than .5 s.d. (in favor of the control group); and this will be taken into account in the statistical analysis.

The WWC could object to this study because of the date, although it was published less than 20 years ago. It could also object because *RM* is not explicitly mentioned or because the percentage of students with learning disabilities is not identified.

11) McIntyre, E., Rightmyer, E. C., & Petrosko, J. P. (2008). Scripted and non-scripted reading instructional models: Effects on the phonics and reading achievement of first-grade struggling readers. *Reading and Writing Quarterly, 24*(4), 377–407.

12) Rightmyer, E. C., McIntyre, E., & Petrosko, J. P. (2006). Instruction, development, and achievement of struggling primary grade readers. *Reading Research and Instruction, 45*, 209–241.

These articles report on the same data set. First grade teachers in 12 to 17 different schools and 37 classrooms were asked to nominate "struggling" students for inclusion in the study. There were two to five students in each class. Gains in reading achievement over time were compared for students in schools using *Reading Mastery* (n = 56 in Grade 1) and those using other models (total n = 52 in Grade 1), all described as "non-scripted": *Breakthrough to Literacy, Early Success, Four Blocks,* and *Together We Can.* Clay's Hearing Sounds in Words was used to measure gains from fall to spring of first grade and the Flynt Cooper Informal Reading Inventory was used to measure gains from the beginning of first to end of second and the beginning of second to the end of third. Pretesting occurred in September and posttesting in May. The 13 effect sizes calculated ranged from –.73 to .88, with an average of .15.

The WWC might object to these articles because they did not specifically mention "learning disabilities." In addition, gain scores, rather than pretest scores, are reported, so it is difficult to ensure that the groups were equivalent at pretest.

13) O'Connor, R. E., Jenkins, J. R., Cole, K. N., & Mills, P. (1993). Two approaches to reading instruction with children with disabilities: Does program design make a difference? *Exceptional Children, 59*(4), 312-323.

This study involved a transitional kindergarten program for children with disabilities. Language, cognitive, motor, and socio-developmental delays are noted, but "learning disabled" was not in the list. Students were randomly assigned to *Reading Mastery* or to the Addison Wesley *Meet the Superkids* and the *Superkids' Club* programs. Pre- and post-intervention measures were administered: the McCarthy Scales of Children's Abilities, Test of Early Reading (TERA), California Achievement Test (CAT), and the Peabody Individual

Achievement Test. The 18 effect sizes ranged from –.21 to +.91, with an average value of .30.

As noted above (section III, study 4), the WWC rejected this study because fewer than half of the students were identified as having LD.

14) Richardson, E., DiBenedetto, B., Christ, A., Press, M., & Winsberg, B. G. (1978). An assessment of two methods for remediating reading deficiencies. *Reading Improvement, 15*(2), 82–95.

Students in Grades 2–6 were recommended for participation by teachers in two schools. Pretests were given to ensure that their reading skills were at least one year behind chronological age. Students in each school were matched for age and PIAT reading scores and divided into two groups. One received instruction in *DISTAR* or *CR* (for the older students) and the other in ISM, a program developed by one of the authors. Teachers were trained. Four months of instruction were given. Increases in skills were found for both groups, with slightly larger (although insignificant) gains for the ISM group. The four calculated effect sizes ranged from –.18 to –.03, with an average of –.10.

The WWC could reject this because of the age of the study, the use of *DISTAR* as well as *CR*, the lack of a reference to learning disabilities, and the use of gain scores with reference to a normative sample without reporting pretest measures.

15–18) SRA/McGraw-Hill case studies

SRA has disseminated a number of "case studies" under the title "Results with *Reading Mastery.*" A recent article (Stockard, 2013) I showed how these data fit the criteria for cohort control designs, the type of design that the Shadish, Cook, Campbell, and Stanley tradition deems especially appropriate for organizational settings such as schools. Because succeeding cohorts within a given school tend to be relatively similar to each other (e.g., this year's first graders are relatively similar to last year's in beginning skills and parental SES), they can serve as comparison groups. In fact, comparisons between such cohort groups are often the question that is of most interest to school officials. Their key concern generally is "How does a new curriculum impact the students in my school?" The data in the SRA case studies do this very nicely by comparing achievement of cohorts without *RM* to that of cohorts with *RM*. Because the cohorts are from the same school it is generally reasonable to assume that the cohorts are quite similar. (And if they are not, the school officials, or at least the teachers, are quite aware of these differences.) Thus a cohort control group design is seen as internally valid. Even more important, this design is externally valid, without the confounding issues that can result when introducing artificial circumstances, such as

random assignment, into intact settings. (The 2013 article expands on these points with extensive citations to the methodological literature.)

Four of the case studies reported by SRA/McGraw-Hill involved the implementation of *RM* with students who might be seen as having learning disabilities. The results of these case studies are described below. The WWC may reject these studies because the schools used both *RM* and *CR*, as appropriate for the grade level of the students. They may also reject the results because there were no pretest data.

15) SRA/McGraw-Hill. (2006a). *Exceptional education and regular education students excel with Direct Instruction*. Retrieved from SRA website: http://www.mheresearch.com/assets/products/c9f0f895fb98ab91/iredell_statesville_schools.pdf

This report describes data from the Iredell Statesville School District in North Carolina. In fall 2003, teachers in the special education department adopted *Reading Mastery* and *Corrective Reading* for intervention in Grades K–12. The percentage of special education students attaining adequate yearly progress changed from 43% in 2002–03, before implementation of *RM* and *CR*, to 66% in the 2005–06 academic year. This represents an effect size of .46.

16) SRA/McGraw-Hill. (2006c). *Reading proficiency more than doubles among Putnam County special education students. http://www.mheresearch.com/assets/products/c9f0f895fb98ab91/putnam_county_schools.pdf*

This report describes data from the Putnam County School District, based in Cookeville, Tennessee. The district began implementing DI programs (*RM* for Grades K–3 and *Corrective Reading* for Grades 4–8) in all special education classrooms in the 2003–04 school year. The report gave the percentage of special education students who read at the proficient level, as measured by the state of Tennessee's assessment program (TCAP), in 2002–03 before implementation, through 2005–06. The associated effect size was 1.14.

17) SRA/McGraw-Hill. (2008a). *Special education students at California elementary school achieve AYP with Direct Instruction. http://www.mheresearch.com/assets/products/c9f0f895fb98ab91/primrose_elementary_school.pdf*

This report provides data on changes over time in the achievement of students at Virginia Primrose School in Fontana, California. After the introduction of DI programs (*RM* in Grades K–3 and *CR* in Grades 4–5), the percentage of special education students achieving

adequate yearly progress (AYP) moved from 21.2% before implementation to 37.2% by the end of the second year of implementation, an effect size of .35.

18) SRA/McGraw-Hill. (2008b). *Direct Instruction reduces special education referrals in Louisiana school district by half*. http://www.mheresearch.com/assets/products/c9f0f895fb98ab91/rapides_school_district.pdf

This report focused on changes in two schools in the Rapides Parish school district in Louisiana. At the start of the 2006–07 school year, the district introduced *Reading Mastery* and *Corrective Reading* due to concerns with high numbers of students being referred for special education. Over time the number of referrals for special education evaluations dropped by 50% and the proportion of students meeting promotional standards increased. Effect sizes regarding changes in the percentage of SPED students meeting promotional standards were calculated for two schools. The effect in one school was .95 and in the other was .20 (average = .58).

19) Stein, C., & Goldman, J. (1980). Beginning reading instruction for children with minimal brain dysfunction. *Journal of Learning Disabilities, 13*(4), 52–55.

Subjects in this study were 63 students at a private school for children with serious learning problems who had been diagnosed with "minimal brain dysfunction of specific learning disability" (p. 53). One group used *DISTAR* and the other used the *Palo Alto Reading Program*, described as a "cognitive" approach. Pretests on the PIAT showed no significant differences between the groups. Posttest results favored the *DISTAR* group (d = 1.35). Somewhat stronger results appeared when the analysis was limited to a smaller subset matched on variables that were correlated with pretest scores (d=2.02).[15]

The WWC could reject this study because of the date of publication and because it used *DISTAR* rather than *RM*. Note, however, that it specifically mentioned learning disabilities. It could also reject the study because it used gain scores and did not report pretest values.

20) Stockard, J. (2008). *Reading achievement in a Direct Instruction school and a "three tier" curriculum school.* Eugene, OR: National Institute for Direct Instruction, Technical Report 2008-5.

This report examines data from two schools within the same Oregon school district. One school used *RM* as the core reading curriculum for all primary children while the other used

[15] The effect sizes (Cohen's d) were calculated from the *t*-ratios (5.27 for the total group and 3.77 for the smaller, matched sample), using the online calculator found at http://easycalculation.com/statistics/effect-size-t-test.php.

a “three tiered” model, occasionally employing DI for students that teachers felt would benefit from the instruction. Data were available for two cohorts of students who were in the schools from kindergarten through third grade. There were almost equal numbers of students from each school and there were no significant differences between the schools in the students’ eligibility for free or reduced lunch, their racial-ethnic characteristics, their special education designation, or their DIBELS scores at the start of kindergarten. Results are given separately for special education (SPED) and general education students. Students in the Direct Instruction school had statistically significantly higher gains in nonsense word fluency (NWF) through the end of first grade and in oral reading fluency (ORF) from first through third grade than students in the control school. These differences were especially marked for students in special education, where the effect size for NWF at the end of first grade was .70 and the effect size for ORF at the end of third grade was .72. Perhaps even more important, at the end of third grade the reading scores of the SPED students in the DI school were not significantly different from those of the general education students in the comparison school (d = –.12 in favor of the control school).

The WWC could reject this study because it did not involve random assignment and because the students are not specifically identified as being learning disabled.

21) Umbach, B., Darch, C., & Halpin, G. (1989). Teaching reading to low-performing first graders in rural schools: A comparison of two instructional approaches. *Journal of Instructional Psychology, 16*(3), 112–121.

Subjects in this study were 31 first grade students nominated by their classroom teachers as students who were “having difficulty with reading and needed extra help” (p. 114). The students were randomly assigned to receive *RM* or the Houghton-Mifflin program. There were no significant differences at pretest between the groups in the Otis-Lennon School Abilities test or the total reading score of the Woodcock Reading Mastery test, and IQs fell within the normal range. The control groups were taught by their regular classroom teachers and the experimental groups by masters degree students from a nearby university. Significant differences between the groups, in favor of the *RM* group, appeared at posttest in both the word identification and passage comprehension subtests of the WRMT. (Posttest scores were adjusted for pretest scores.) No standard deviation was given for the word attack subtest and the effect size was calculated based on the given *t* value of .86, df = 29. Results were also given for the total reading score and the statistical analysis described below adjusted for this inclusion. The four effect sizes ranged from .32 to 3.94, with an average of 1.68.

The WWC would probably reject this article because of the date and because learning disability is not explicitly mentioned.

To summarize, this section describes 21 studies that the WWC could have used to examine the efficacy of *Reading Mastery* with students who have some type of learning disability. Brief descriptions of the methods and results were given as well as speculations regarding why the studies would not meet the WWC's criteria for review. The next section looks at the studies in a more systematic, statistical fashion, attempting to see if variations in the studies' designs and, especially, variations in the ways they do or do not meet the WWC's criteria, are related to variations in effect size.

B. Brief Statistical Analysis of Associated Effect Sizes

This section uses standard statistical techniques to analyze variations in the effect sizes associated with the findings of the studies listed in the above section. First, descriptive data on the effects and study characteristics are provided, summarizing the information presented above. Then results of a mixed model analysis, which looks at the ways in which the characteristics of the studies are related to the effect sizes, are given. In other words, the analysis looks at the impact of the WWC criteria on effect sizes within the studies. One could suggest that if the WWC criteria are important to include, they should be related to variations in effect sizes in a systematic and statistically significant manner. Because the literature regarding the efficacy of DI programs is so large, it is especially well suited to address this question.

B.1 Summary descriptive analyses

Table 1 summarizes data on each of the 21 studies included in the meta-analysis and described above.[16] The eight columns following the citation indicate if a study met each of the selection criteria that the WWC uses. The final column reports the total number of criteria that each study met (termed "Total score"). The bottom row reports the number of studies that met each criterion. The columns are ordered in terms of most frequent occurrence. More than three quarters of the studies used a design other than norm comparison, occurred in the U.S., and used only achievement-related factors in the student selection process. More than half reported pretest data and were published in the last 20 years. Substantially fewer (8 of the 20) employed random assignment of students or teachers. Fewer still (6 of the 20) used only *Reading Mastery* (and not *DISTAR* as in the older studies or adding *Corrective Reading*, as recommended, for the older students). Only 2 of the 20 studies included an explicit statement indicating that more than 50% of the subjects had learning disabilities. None of the studies met all eight of the criteria listed. One

[16] Because they report on the same data set, the data for McIntyre, Rightmyer, & Petrosko (2008) and Rightmyer, McIntyre, & Petrosko (2006) are combined (making the total number of studies 20 rather than 21). The three studies by Gunn and associates are separated in Tables 1 and 2 because they report on data from different years. However, in the mixed model analysis they are treated as one study (level 2 grouping), a decision that is more conservative in nature (limiting degrees of freedom and thus the probability of significant results in favor of DI).

Table 1

Characteristics of Studies Included in the Analysis

Study	Norm Comp.	in U.S.	No other selection factors	Recent	Pretest	Random	*RM*	50% LD	Total score
Bowers (1972)	Yes	Yes	Yes	No	Yes	Yes	No	No	5
Branwhite (1983)	Yes	No	Yes	No	No	No	No	No	3
Cooke et al. (2004)	No	Yes	Yes	Yes	Yes	No	Yes	Yes	6
Francis 1991	Yes	No	Yes	No	Yes	No	Yes	No	4
Gersten & Maggs (1982)	No	No	Yes	No	Yes	No	No	No	2
Gunn et al. (2000)	Yes	Yes	No	Yes	Yes	Yes	No	No	5
Gunn et al. (2002)	Yes	Yes	No	Yes	Yes	Yes	No	No	5
Gunn et al. (2005)	Yes	Yes	No	Yes	Yes	Yes	No	No	5
Kamps et al. (2003)	Yes	Yes	No	Yes	No	No	No	No	3
Marston et al. (1995)	Yes	Yes	Yes	Yes	Yes	Yes	No	No	6
Rightmyer/McIntyre, & Petrosko (2006 & 2008)	Yes	Yes	Yes	Yes	No	No	Yes	No	5
O'Connor et al. (1993)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	7
Richardson et al. (1978)	No	Yes	Yes	No	No	Yes	No	No	3
SRA/McGraw-Hill (2006a)	Yes	Yes	Yes	Yes	No	No	No	No	4
SRA/McGraw-Hill (2006c)	Yes	Yes	Yes	Yes	No	No	No	No	4
SRA/McGraw-Hill (2008a)	Yes	Yes	Yes	Yes	No	No	No	No	4
SRA/McGraw-Hill (2008b)	Yes	Yes	Yes	Yes	No	No	No	No	4
Stein & Goldman (1980)	Yes	Yes	Yes	No	No	No	No	Yes	4
Stockard (2008)	Yes	Yes	Yes	Yes	Yes	No	Yes	No	6
Umbach et al. (1989)	Yes	Yes	Yes	No	Yes	Yes	Yes	No	6
Total "Yes" Codes (out of 20)	17	17	16	13	11	8	6	2	

Note: Articles were coded as "recent" if they had appeared in the last 20 years (1993–2012). Random assignment refers to random assignment of either students or teachers. The *RM* code indicates that only *Reading Mastery* and not *DISTAR* or *RM* plus *Corrective Reading* was used. No other factor in selection indicates that other criteria, such as behavior, were not part of the student referral process.

(O'Connor et al., 1993) met 7 of the 8 criteria; and four others met 6 of the 8 criteria. Four studies met only two or three of the criteria.[17]

Table 2 summarizes the information on effect sizes associated with each study described in the previous section. Across the 20 studies, the smallest average effect size reported was –.48, while the largest was 1.71 (obtained from the column labeled "Average") The average effect size across all the studies was .53, more than twice the level of .25 generally used to denote educational importance.

Table 2

Effect Sizes and Comparison Counts of Studies in Meta-Analysis

		Effect sizes			
	Study	Min.	Max.	Average	Count
1)	Bowers (1972)	-0.39	0.60	0.15	4
2)	Branwhite (1983)	1.71	1.71	1.71	1
3)	Cooke et al. (2004)	0.20	0.71	0.34	6
4)	Francis 1991	-0.70	0.89	0.11	20
5)	Gersten & Maggs (1982)	0.36	1.08	0.72	2
6)	Gunn et al. (2000)	0.00	0.79	0.34	24
7)	Gunn et al. (2002)	0.05	0.74	0.31	15
8)	Gunn et al. (2005)	0.03	0.40	0.23	10
9)	Kamps et al. (2003)	0.19	0.73	0.39	6
10)	Marston et al. (1995)	-0.84	0.01	-0.48	6
11)	McIntyre, et al. (2008)*	-----	-----	-----	-----
12)	Rightmyer, et al (2006)	-0.73	0.88	0.15	13
13)	O'Connor et al. (1993).	-0.21	0.91	0.30	18
14)	Richardson et al. (1978)	-0.18	0.03	–0.10	4
15)	SRA/McGraw-Hill. (2006a)	0.29	0.29	0.29	1
16)	SRA/McGraw-Hill. (2006c)	1.14	1.14	1.14	1
17)	SRA/McGraw-Hill (2008a)	0.35	0.35	0.35	1
18)	SRA/McGraw-Hill (2008b)	0.20	0.95	0.58	2
19)	Stein & Goldman (1980)	1.35	2.02	1.69	2
20)	Stockard (2008)	0.70	0.72	0.71	2
21)	Umbach et al. (1989)	0.32	3.94	1.68	4
	Total	-0.84	3.94	0.53	20

*As explained in the text, the Rightmyer, et al (2005) and McIntyre, et al (2008) studies use the same data, so only one entry is given for those citations.

The top rows of Table 3 give descriptive statistics on all the effect sizes (n = 146) used in the analysis. In other words, while Table 2 aggregates effect sizes to the average for each

[17] The rolling 20-year criterion will, however, soon result in O'Connor not meeting the criterion of publication in the last 20 years, resulting in the article meeting only 6 of the 8 criteria.

study, the data reported in Table 3 use each comparison as the unit of analysis. This results in more weight given to some studies (those with more comparisons) than to others and also introduces substantially more variability to the results. The average effect size across these comparisons is .31, somewhat lower than in Table 2. When a weighted effect size is used (adjusting for the number of cases in each comparison), the average effect size is .30, virtually identical to the unweighted value.

Table 3

Descriptive Statistics for Variables in the Meta-Analysis (Comparison as Unit of Analysis)

	Mean	SD	Min.	Max.
Comparison Level Variables				
Effect size (Cohen's d)	0.31	0.52	–0.84	3.94
Number of effect sizes	7.33	5.83	1	24
Sample size of comparison	76.5	58.4	12	275
Sample size estimated (1 = yes)	0.06	0.24	0	1
Aggregate of others in the list (1 = yes)	0.19	0.4	0	1
Posttest means adjusted (1 = yes)	0.38	0.49	0	1
Pretest scores > .5 s.d. (1 = yes, n = 46)	0.2	0.4	0	1
Grade in comparison (k = 0)	2.23	1.45	0	6
Grade estimated based on age	0.02	0.14	0	1
Multiple years in grade estimate (number of extra years)	1.87	1.9	0	5
Study Level Variables				
Year of publication	1997.36	8.53	1972	2008
Published after 1993 (less than 20 years ago)	0.75	0.44	0	1
Pretest data given	0.79	0.41	0	1
Random assignment students or teachers	0.58	0.49	0	1
50% of students with LD explicit	0.05	0.23	0	1
Reading Mastery only	0.43	0.5	0	1
No other factor in selection	0.62	0.49	0	1
Not norm comparison	0.92	0.28	0	1
In U.S.	0.84	0.37	0	1
Total design score	4.99	1.1	2	7

Note: The unit of analysis is the comparison (equal to the count variable in Table 2). There were 146 comparisons for all of the variables excepting the comparison of pretest scores, where data were available for only 46 of the comparisons.

The other entries in the top part of Table 3 provide additional information about the comparisons. On average, there were slightly more than 7 effect sizes associated with each study, and an average of 76 cases per comparison. The effect sizes were aggregates of others in the analysis in about 20% of the cases and the posttest means were adjusted (usually for pretest scores) in over a third of the cases. In a few cases (20% of those with pretest data provided and 6% of the total group), pretest scores varied by more than .5 of a standard deviation. (In most cases the differences involved the control group scoring higher at pretest, a common result when the intervention is targeted at the most needy students.) Results were given for students from kindergarten through sixth grade, with the average comparison reflecting second graders. Grade level was estimated from the age of the students in only 2% of the cases, and data were given for a range of grades in slightly more than half of the comparisons, with an average of almost 2 extra years (3 years in total).

The bottom rows of Table 3 give descriptive statistics for the study level variables. The proportions differ from those in Table 1 because some studies have more effect sizes than others, although the results are substantively similar. The majority of effect sizes came from studies published in the last 20 years, included pretest data, were conducted in the U.S., and did not involve a norm comparison design. Over half involved random assignment of students and teachers and had no factor (such as behavioral problems) used in the sample selection process. Slightly fewer than half used only *Reading Mastery*. Others used *DISTAR* or added *Corrective Reading* for older students. Finally, only 5% of the effects came from studies that explicitly stated that 50% or more of the students had learning disabilities.

In short, while none of the studies would appear to meet all of the criteria that the WWC has established for studies that it will review, they have a fair amount of variability in these criteria. While the average effect size is positive, there is also variation in effects from one study to another. The next section uses statistical techniques to see if the variations in study characteristics are related to variations in effect size.

B.2 Mixed model analyses

Mixed models were used to examine the way in which study characteristics are related to effect size. Mixed models are simply an extension of linear regression, but are especially useful when one has data on two or more levels—in this case multiple effect sizes in a set of studies.[18] For this analysis the study characteristics are called "level 2" and the effect sizes within each study are called "level 1." Explanatory variables at level 2 are the criteria that the WWC uses to select studies for review (in the bottom rows of Table 3 and listed in Table 1), while the explanatory variables at level 1 are the characteristics associated with the effect sizes and listed in the top rows of Table 3.

[18] Mixed models are actually the regression equivalent of a nested factor analysis model.

The mixed models were conducted with a variety of restrictions on the sample and the results were virtually identical to those presented here. The results obtained with the most conservative sample are presented below. For this analysis the four most extreme effect size values (two from the bottom and two from the top of the distribution) were omitted. As would be expected, the average value of the effect size was altered relatively little, while there was a much more marked impact on the standard deviation.[19] The three studies by Gunn and associates were grouped together (treating as one level 2 study). Finally, any effect sizes that were the aggregates of others in the data set and cases where pretest differences were more than .5 of a standard deviation were omitted. This resulted in a sample of 109 effect sizes from 17 studies.[20] The studies included had 1 to 36 effect sizes, with a mean number of effects of 6.4 per study.[21] The average effect size with this sample (using comparisons or the level 1 values as the unit of analysis) was .30 (s.d. = .38, minimum = –.73, maximum = 1.71).

Table 4 summarizes the results. Each line in Table 4 reports the results from a separate analysis. The first is a simple intercept-only model, having each study as a random effect (the first line of data in the table). This is equivalent to an analysis of variance with the studies as a factor and tests the null hypothesis that the effect sizes are equivalent across studies. The intercept in the resulting equation is equivalent to the average effect size across the studies but adjusts for the multiple numbers of effects for each study. Note that the value of .44 is somewhat greater than the simple mean of .30 obtained without introducing the study level controls.

The next model adds the sample size of each comparison to the equation, and these results are shown in the second line of Table 4. Although sample size had no significant relationship with effect size, this control resulted in a slight lowering of the intercept. In other words, with sample size equalized and adjusting for the multiple entries in each study (as in the second line of data), the average effect size is .37.

The following entries in the table all involve equations with two variables: sample size and one of the comparison or study level variables listed in Table 3.[22] For instance, the third line of data gives results of the analysis including grade of students and sample size, and the

[19] The average effect size in the reduced sample was .29, s.d. = .38, min. = –.73, max = 1.71 (n = 142). As shown in Table 3, the average effect size in the full sample was .31, s.d. = .52, min. = –.84, max = 3.94 (n = 146). If anything, the omission of outliers had a conservative impact on the results by diminishing the reported effect size.

[20] Results from Stein & Goldman (1980) were omitted from this reduced sample because the effect sizes were outliers.

[21] A variety of combinations of variables in predictive equations were also examined, and the results were all substantively similar to those reported here.

[22] Only two predictor variables were included in each equation to maximize the degrees of freedom. There was no indication that the substantive nature of the results altered when multiple predictor variables were used.

fourth line gives the results with the indicator of adjusted posttest means and sample size, and so forth. The coefficients and probabilities associated with the intercept (the effect size) are in the first two columns of data. The values associated with the study characteristics are in the third and fourth columns of data, and those associated with sample size are in the final columns. The fixed effect (akin to a regression coefficient) is given in the column labeled "effect" and the associated probability is given in the column labeled "Probability."[23]

Table 4

Mixed Model Results Regressing Effect Size on Sample Size, Comparison Level, and Study Level Variables

	Intercept		Variable		Sample Size	
Variable	Effect	Prob.	Effect	Prob.	Eff.*10^2	Prob.
Intercept only	0.44	<.001	-----	-----	-----	-----
Sample size of comparison	0.37	0.004	-----	-----	0.14	0.30
Comparison Level (Level 1) Variables						
Grade in comparison	0.61	<.001	–0.08	0.04	0.10	0.44
Post–test means adjusted	0.40	0.004	–0.07	0.72	0.14	0.30
Study Level (Level 2) Variables						
Recently published	0.53	0.004	–0.26	0.24	0.15	0.25
Pretest data given	0.52	0.006	–0.21	0.32	0.10	0.43
Random assignment	0.47	0.001	–0.26	0.22	0.13	0.32
50% LD explicit	0.38	0.005	–0.06	0.89	0.13	0.31
Reading Mastery only	0.36	0.026	0.02	0.93	0.14	0.30
In U.S.	0.71	0.005	–0.43	0.12	0.16	0.21
Total design score	0.83	0.028	–0.10	0.2	0.09	0.46

Note. The measure of no other factor in design selection was omitted because it was so highly correlated (near .70) with the sample size. The measure of norm comparison designs was omitted because it was so highly skewed (10/90 split). To obtain the actual fixed effect associated with the sample size, multiply the given coefficient by 10^2 (100).

The fixed effects associated with the intercept are always positive and statistically significant. In other words, no matter what type of study level or comparison level control variable was used, the intercept was positive and significantly greater than zero. In addition, all of the coefficients associated with the intercept are greater than .25, often substantially so. Recall that this is the level often used to denote educational importance. Thus, there is

[23] In an attempt to keep the results simple a number of details about the analyses have been omitted, notably the standard errors associated with the fixed effects and various model fit statistics. These details can be provided upon request.

no reason from this analysis to suspect that the factors the WWC uses for selection criteria would significantly diminish the estimated efficacy of the DI programs. In fact, the average coefficient associated with the intercept across the seven analyses incorporating study level (level 2) characteristics is .54, an increase of over 20% from the level in the intercept only or baseline model.

The other coefficients in Table 4 examine the impact of sample size and each of the study and comparison level variables. Only one of these coefficients is significant at the .05 level, a result that would be expected by chance. Most notably, none of the study level variables (those in the last seven lines) are significantly related to effect size. Most of the coefficients associated with the study characteristics are negative, indicating that effect sizes are somewhat smaller when these characteristics are present, but none is significantly different from zero—either separately or in the composite (the last line of data). In other words, none of the criteria that the WWC uses to screen studies were significantly related to variations in effect size. This could suggest that restricting studies for review by these characteristics does not make any difference in the results.

V. Summary

This report provides an analysis of why the conclusions of the What Works Clearinghouse regarding Direct Instruction differ so markedly from the extant scholarly literature. The first section describes a number of procedural characteristics of the reviews that may contribute to this issue. The problems described involve both their criteria regarding exclusion or inclusion of studies for review and the ways in which these reviews are apparently conducted. Key issues noted include the ways in which reviews focus on narrow curricular programs, fail to examine or consider the characteristics of the programs, apply an arbitrary time limit to the included studies, and use "standards" for review that differ markedly from those used in the general social science world and result in a seemingly blanket exclusion of field-based studies and those using advanced statistical methods.

While the first section focuses on general theoretical and methodological issues, the second section examines the 10 studies of DI programs that the WWC has found to meet their inclusion criteria, either with or without reservation. Serious errors in the decisions regarding 4 of the 7 studies that were deemed as meeting their criteria "without reservation" are documented. In other words, of the studies that were deemed to meet their criteria fully, more than half probably should not have been included for review because they do not provide an appropriate test of the programs.

The third and fourth sections turn to the content of the WWC report on *Reading Mastery* and students with learning disabilities that was initially posted in July 2012. The third section examines each of the studies included in the report and provides an assessment regarding

the decision to include or exclude the study from review. Of the 17 studies listed in the report, the classification of 4 was questioned.

The fourth section continues the focus on the use of *Reading Mastery* with students with learning disabilities. The first sub-section describes 21 research studies that could have been included in the WWC review, detailing the design and conclusions of the studies, the effect size associated with their results, and reasons that the WWC might reject the study for inclusion. The second sub-section reports a statistical analysis of the effect sizes, specifically testing the extent to which meeting a WWC criteria is related to effect sizes. In other words, this analysis tests the hypothesis that the criteria used by the WWC provide a more accurate estimate of the efficacy of a curriculum. None of the study criteria used for inclusion/exclusion by the WWC was significantly related to the magnitude of effect sizes. In other words, the criteria used by the WWC are not related in any significant or systematic way to the impact reported in a study. The criteria used to eliminate studies from consideration appear to have no systematic or significant relationship to the direction or magnitude of the results that were reported.

The notion of science as a cumulative enterprise has long dominated scholarly thinking. When science is viewed as a cumulative enterprise, scholars examine results over a long period of time, in varied settings, and with different populations. They assume that results will "balance out," and that, over time, as findings accumulate across a range of settings and populations, similarities and variations in results will become apparent and we will have better knowledge of the actual state of affairs. The method of meta-analyses and the tradition of "generalized causal inference" touted by Shadish, Cook, and Campbell (2002) are representative of this long accepted approach (see Stockard, 2013 for an extended discussion).

The very restrictive approach that the WWC has taken to its consideration of studies for review is a sharp departure from this tradition, and it appears that this departure can account for many of the discrepancies between the established literature regarding DI and the WWC's conclusions. The WWC has opted to base its conclusions on the results of studies that conform to a detailed set of methodological criteria. As described above, this approach is so restrictive that it results in very few studies being accepted for analysis. As a result, any conclusions are based on a very small sub-sample of the available evidence, and many high quality studies, especially those using large samples and advanced statistical techniques, appear to be excluded.

Because the WWC conclusions are based on such a small sample, it is crucial that their analysis be error free and that the criteria employed lead to more accurate results. The data presented above suggest that this has probably not occurred. First, serious errors in decisions about both the inclusion and exclusion of studies were described. Many studies

were not even reviewed for possible inclusion, and the decisions for rejection or acceptance of those that were reviewed were often questionable. Second, the interpretations of the few studies that were accepted can sometimes be challenged, as exemplified by the analysis of the Cooke et al. 2004 study described above. Finally, there is no indication that the criteria used to select or exclude studies from consideration are related to the results that are obtained. Estimates of the efficacy of DI programs are equivalent regardless of the nature of the WWC criteria that are examined. One could logically conclude that a wider net, which accepted more studies for review, would present a much more accurate picture of the research results.

In short, the WWC procedures appear to result in a very selective and inaccurate view of the DI literature, and this is probably the basis for the discrepancy between their conclusions and those of the research literature. Consumers would be well advised to consult sources of summary material other than the WWC and, especially, the well-conducted and highly regarded meta-analysis literature (e.g., Adams & Engelmann, 1995; Borman, Hewes, Overman, & Brown, 2003; Hattie, 2009; and White, 1988).

References

Adams, G. L., & Engelmann, S. (1995). *Research on Direct Instruction: 25 years beyond DISTAR.* Seattle, WA: Educational Achievement Systems.

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*(2), 125–230.

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

Carlson, C. D., & Francis, D. J. (2002). Increasing the reading achievement of at-risk children through Direct Instruction: Evaluation of the Rodeo Institute for Teacher Excellence (RITE). *Journal of Education for Students Placed At Risk, 7*(2), 141–166.

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

Good, R. H., Wallin, J., Simmons, D. C., Kame'enui, E. J., & Kaminski, R. A. (2002). System-wide percentile ranks for DIBELS Benchmark Assessment (Technical Report 9). Eugene, OR: University of Oregon.

Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education, 39*(2), 66–85

Hattie, John A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.

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

Stockard, J. (2008). *The What Works Clearinghouse beginning reading reports and rating of* Reading Mastery*: An evaluation and comment.* Eugene, OR: National Institute for Direct Instruction, Technical Report 2008-4.

Stockard, J. (2010). An analysis of the fidelity implementation policies of the What Works Clearinghouse. *Current Issues in Education, 13*(4).

Stockard, J. (2013). Merging the accountability and scientific research requirements of the No Child Left Behind Act: Using cohort control groups," *Quality and Quantity: International Journal of Methodology*, 47, 2225-2257, available online, December 2011.

Stockard, J. (2012). *A summary of concerns regarding the What Works Clearinghouse.* Eugene, OR: National Institute for Direct Instruction.

Stockard, J., & Wood, T. W. (2012). Reading Mastery *and learning disabled students: A comment on the What Works Clearinghouse Review.* Eugene, OR: National Institute for Direct Instruction.

Tallmadge, G. (1977). *The joint dissemination review panel idea book*. Washington, DC: NIE, U.S. Government Printing Office.

Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., et al. (2006). *National assessment of title I. Interim report. Volume II. Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., et al. (2007). *National assessment of Title I. Final report. Volume II. Closing the reading gap: Findings from a randomized trial of four reading interventions for striving readers (NCEE 2008-4013)*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

What Works Clearinghouse. (2006). *Intevention Report,English Language Learners. Reading Mastery/SRA/McGraw Hill.* Washington, DC: Institute of Education Sciences

What Works Clearinghouse. (2007a). *Intevention Report,Early Childhood Education/Direct Instruction, DISTAR, and Language for Learning.* Washington, DC: Institute of Education Sciences

What Works Clearinghouse. (2007b). *Intevention Report, Beginning Reading/Corrective Reading.* Washington, DC: Institute of Education Sciences

What Works Clearinghouse. (2008). *Intevention Report, Beginning Reading/Reading Mastery.* Washington, DC: Institute of Education Sciences

What Works Clearinghouse. (2010a). *Intervention Report, Adolescent Literacy/Corrective Reading.* Washington, DC: Institute of Education Sciences.

What Works Clearinghouse. (2010b). *Intervention Report, Adolescent Literacy/Reading Mastery.* Washington, DC: Institute of Education Sciences

What Works Clearinghouse. (2012). *Intervention Report, Students with Learning Disabilities /Reading Mastery.* Washington, DC: Institute of Education Sciences

White, W. A. T. (1988). A meta-analysis of the effects of Direct Instruction in special education, *Education and Treatment of Children, 11*(4), 364–374.