National Institute for Direct Instruction

Does the What Works Clearinghouse Work?[a]

A NIFDI White Paper

Jean Stockard
National Institute for Direct Instruction and University of Oregon
and
Timothy W. Wood
National Institute for Direct Instruction

September 9, 2013

a) We thank Jerry Silbert for very helpful comments on earlier drafts of this paper. Any errors in the manuscript and any opinions expressed are the sole responsibility of the authors.

Does the What Works Clearinghouse Work?

Abstract

The What Works Clearinghouse (WWC) is a federally funded program established in 2002 to evaluate educational interventions and provide reliable and trustworthy summary ratings and reports of their effectiveness. Yet, some of their reports directly contradict the conclusions of the research literature, giving positive ratings to a program that scholars have found to be ineffective (*Reading Recovery*) and failing to give positive ratings to programs the research literature has found to be highly effective (Direct Instruction). This article uses a comparative case study approach to examine how these contradictory conclusions developed. We contrast the methods used by the scholarly world and the WWC to summarize literature and their conclusions about the two curricula. We then examine errors in the three major steps of the WWC review process: 1) compiling lists of studies to examine, 2) applying WWC criteria to select studies for further analysis, and 3) interpreting and reporting the results of the studies. Extensive problems are documented at each step, systematically favoring *RR* and not favoring DI. Implications of the results are briefly discussed.

Does the What Works Clearinghouse Work?

The What Works Clearinghouse (WWC) is a federally funded program established in 2002 to evaluate educational interventions and provide summary ratings and reports of their effectiveness. The WWC's website describes their organization as a "trusted source of scientific evidence for what works in education to improve student outcomes" and as providing "accurate information on education research" (WWC, 2013a). Several WWC reports, however, raise troubling questions about the accuracy of their work, for the conclusions directly contradict those within the research literature. Specifically, the WWC has given positive ratings to *Reading Recovery*, a program scholars have found to be ineffective, while failing to give positive ratings to Direct Instruction programs that the research literature has found to be highly effective.

This article examines these highly contradictory conclusions using a comparative case study approach. We begin by contrasting the methods used by the scholarly world and the WWC and their conclusions about the curricula. We then examine errors in three major steps of the WWC review process: 1) compiling lists of studies to examine, 2) applying WWC criteria to select studies for further analysis, and 3) interpreting and reporting the results of the studies. We end with a brief discussion of implications of our results.

Two Approaches to Reviews

The notion of science as a cumulative enterprise has long dominated scholarly thinking. Methodologists stress the ways in which scientific knowledge gradually accumulates. They also note that there can be no "perfect" experiment and the importance of looking at a variety of results in a range of settings. As Cook and Campbell put it, "we stress the need for *many* tests to determine whether a causal proposition has or has not withstood falsification; such determinations cannot be made on one or two failures to achieve predicted results" (1979, p. 31, emphasis in original). Meta-analyses and the tradition of "generalized causal inference" discussed by Shadish, Cook, and Campbell (2002) are representative of this long accepted approach (Stockard, 2013b). The statistical versions translate results into effect sizes, reported on a continuous scale, and include controls for methodological elements that might be seen as influencing the outcomes.

The WWC uses a different process, restricting its analyses to studies that conform to a detailed set of methodological criteria. Like reviews in the scholarly arena, the WWC procedure begins by amassing lists of articles that assess the efficacy of a given curriculum. They then, however, restrict their analysis to studies that conform to an extensive series of requirements such as the nature of the design, with a strong preference for randomized assignment; the date of publication, generally excluding those published more than 20 years prior to the review; and detailed data on areas such as attrition of subjects and pretest scores. Most of their reviews are focused on specific curricula such as named reading series and students in different grade levels or language learning or disability status. Studies that have passed this screen are then analyzed with a set of pre-determined criteria and reported using five summary ratings: positive, potentially positive, mixed, potentially

negative, and negative (WWC, 2013d). Thus, in contrast to the typical scholarly approach, the WWC examines a highly selective segment of the literature, uses selection criteria to control for methodological variations and limit the set of studies examined, and reports results in categorical rather than continuous or interval measures.[1]

WWC and Academic Reviews of Two Literacy Curricula

Our analysis focuses on two reading curricula that represent different approaches to the teaching of literacy skills and that have received strikingly different evaluations by the academic literature and the WWC.

Reading Recovery

Reading Recovery (RR) is a one-on-one tutoring program, based on a constructivist approach, individually tailored, and designed to help first grade struggling readers catch up with their peers. The general consensus of the scholarly community, based on a large number of peer-reviewed studies and research syntheses, is that "there is little evidence to show that *Reading Recovery* has proved successful" (Baker, et al., 2002, p. 1; also Reynolds & Wheldall, 2007; Tunmer, Chapman, Greaney, Prochnow, & Arrow, 2013).

The WWC's analyses of *RR* contrast sharply with these scholarly reviews. It has issued two reports of the use of *RR* with beginning readers (WWC 2008b, 2013c). Five studies met the criteria for review in 2008, but only three met the criteria in 2013. The WWC concluded in 2008 that *RR* had positive effects on general reading achievement and on alphabetics and "potentially positive effects" in other areas. In 2013 it reported positive effects on general reading and potentially positive effects in all other areas. Of the eight ratings reported in the two reports, three were fully positive and five were "potentially positive."

Direct Instruction

The Direct Instruction (DI) curricula, developed by Siegfried Engelmann and his colleagues, are a set of highly structured programs designed to help students learn effectively and efficiently. A large number of studies have examined the extent to which the curricula promote student achievement, and meta-analyses and narrative reviews of these works have found strong evidence of the programs' effectiveness. Hattie (2009) summarized the results of four meta-analyses that included DI, incorporating 304 studies, 597 effects and over 42,000 students. He found an average effect size of .59, with similar results across various student populations and subject matters. Direct Instruction was the only curricular approach with such strong support (see also Borman, Hewes, Overman, & Brown, 2003; White, 1988).

The WWC's reports on DI contrast sharply with the scholarly literature. Since 2006, the WWC has published eight reviews of Direct Instruction curricula (See Table 1). Out of over 200 studies examined, only five fully met the WWC criteria for review and an additional

[1] The WWC bases its categorizations on effect size numbers like those used in the scholarly literature. However, their reports focus on the categories rather than the effect sizes.

three met the criteria "with reservation."[2] From these eight studies the WWC calculated a total of 19 effects (ranging from two to four per review). Almost two-thirds of these (n=12) were characterized as "no discernible effects," four were "potentially positive," and none were "positive."[3]

Clearly, the conclusions of the WWC and the scholarly community differ in regard to both curricula. With *RR*, the differences result in the WWC giving a positive rating to a program judged ineffective by the scholarly literature. With DI, the differences result in the WWC giving a negative rating to a program judged highly effective by the scholarly literature. All of the eight WWC judgments of *RR* were positive, while only four of the 19 judgments of DI were positive. A simple statistical test indicates that the probability that these different ratings resulted by chance is very small.[4]

We turn now to an examination of why and how these discrepancies occurred, looking at three stages of the WWC review process: 1) searching for relevant studies, 2) deciding which studies to review, and 3) interpreting study results.

Compiling Literature to Review

The WWC asserts that its reports provide "comprehensive coverage of the relevant literature" (2013d, p. vi). Their procedure calls for a systematic and comprehensive search for relevant literature using well specified search terms and a wide range of available databases, websites, and other sources. Their procedures and policy handbook reports that "studies are gathered through a comprehensive search of published and unpublished publicly available research literature, including submissions from intervention distributors/developers, researchers, and the public to the WWC Help Desk" (2013d, p.7).

Our analyses indicate, however, that these searches, at least for studies of Direct Instruction, have been incomplete. An extensive analysis of the WWC's review of the DI program *Reading Mastery (RM)* for beginning reading (WWC, 2008a) reported that the WWC's list of retrieved materials omitted almost 100 research articles, all of which could have been found through searching the reference lists of published literature reviews or standard databases.[5] At least some of these works would clearly have met the WWC's review criteria. Some of the omissions involved large, federally funded work, citations that should have been readily available. In addition, well over half of the works listed as being reviewed were not efficacy studies of *RM*, but other types of reports such as testimonials of school officials, content analyses of reading material or studies of other curricula or students outside the selected age range. In short, the WWC's report of the number of

---

[2] Some studies were listed as reviewed in more than one report; two studies were accepted for review in more than one report.

[3] The WWC reported insufficient data to calculate an effect on 6 additional dimensions.

[4] We used Fisher's Exact Test to determine if the distribution of positive and non-positive ratings across the two programs could occur by chance. The resulting probability was .00006.

[5] One of the meta-analyses used to generate the list of studies that the WWC omitted was listed by the WWC as a source they consulted and found "ineligible for review." Apparently the WWC did not follow up on the references included in this meta-analysis, even though such follow-up is standard scholarly practice.

studies gave a misleading representation of the amount of efficacy literature retrieved and considered (Stockard, 2008).

A more recent example of these errors comes from comparing the 2012 and 2013 reports on the use of *RM* with students with Learning Disabilities (LD) (WWC, 2012, 2013b). Like the 2008 analysis, the 2012 report omitted a large proportion of the relevant studies. We sent a full list of potentially relevant studies to the WWC in summer 2012 and again in winter 2013 providing extensive information on work that was missing from the 2012 review. The WWC acknowledged receipt of both documents.[6]

A revised report on the use of *RM* with students with Learning Disabilities was issued in July 2013.[7] Despite having received detailed information on articles that should have been considered, there were only a few additions. Of the six meta-analyses and literature reviews we recommended they consult, only two were added. Of the 18 efficacy articles recommended, only two were added. One of the added articles was the only one that had an average effect size that was negative and large enough to be considered educationally important. In other words, the WWC ignored over four-fifths of the additional relevant material and appears to have "cherry picked" the only study with a substantial negative effect size.

Analysis of the reports on *Reading Recovery* indicates that the WWC may have been much more responsive to initial omissions in their compilation of studies of this program. The 2008 WWC analysis of *RR* identified approximately 100 studies, while the 2013 report identified about twice that many. Of the approximately 100 studies that were added to the 2013 *RR* review, almost half were published before 2008, suggesting that their initial search of the literature was incomplete. The amount of data examined for the 2013 review of *RR* was twice as large as that examined in the 2008 review, an increase of 100 percent. In contrast, the increase in the number of studies examined for the reviews of *RM* for students with learning disabilities was only 29% (from 17 to 22).

Summary

In short, these data indicate that one reason the conclusions of the scholarly literature and the WWC differ involves different results of the search process. Data from reviews of both programs show that the initial literature searches were far from complete but that the extent to which these problems were rectified varied dramatically across the two programs. More disturbing, when given extensive information on relevant literature about DI programs the WWC appears to have used only a very small portion of this material and in a selective manner.

---

[6] For instance, the coordinator in the WWC Learning Disabilities Review acknowledged receipt of the second report in February, 2013, noting that "it was very helpful in our review of *Reading Mastery*." The first communication (Stockard and Wood, 2012) included an extensive list of efficacy studies that had been omitted and citations of meta-analyses and literature reviews that should have been consulted. The second communication (Stockard, 2013a) included a detailed analysis, using meta-analytic techniques, of 21 relevant studies.

[7] As noted in the footnote to Table 1, the 2013 report was dated 2012. No mention was made of the changes from 2012, even though, as described more below and in Table 1, the reports differed substantially.

Choosing Studies for Review

Once a list of studies has been developed, the WWC reviews each one to determine if it meets criteria for further examination. Our analysis of these decisions indicates serious errors in the application of these criteria that, again, have resulted in biases for *RR* and against DI.

Confounds in Design

One of the WWC selection criteria involves the nature of the research design and, specifically, the extent to which reported results reflect the intervention rather than other factors.[8] Two examples of misapplication of the WWC criteria in this area involve studies of Direct Instruction programs. In both cases, the WWC claimed that there were flaws in the design, while a closer examination of the studies leads to the opposite conclusion.

The WWC used this criterion to reject a report of a large-scale field study (Carlson & Francis, 2002) for its review of *RM* for beginning readers. The study involved thousands of students, hundreds of teachers, and dozens of schools and used sophisticated multivariate techniques. The only intervention in the experimental schools was *RM*, and as with other studies, the results strongly favored the curriculum. In describing the program, Carlson and Francis noted that teacher training in instructional techniques and behavior management had been provided. The incorporation of positive reinforcement and classroom management is integral to DI programs and well documented in all writings about the curriculum and discussed extensively in the teachers' manuals. Yet, the WWC chose to interpret the mention of behavior management as a confounding influence and used this as a reason to exclude the study from consideration (see Stockard, 2008, for supporting details).

A second case involves a large scale, federally funded multi-year study that incorporated random assignment of students to treatment, a large variety of assessments, and both a behavioral intervention and *Reading Mastery* (Gunn, Biglan, Smolkowski, & Ary, 2000; Gunn, Smokowski, Biglan, & Black, 2002). The articles were accepted without reservation for the 2006 WWC analysis of Direct Instruction programs for English Language Learners, but were rejected from the 2008 analysis of *RM* for beginning readers because of a supposed confound.[9] If the articles were truly ineligible under WWC criteria, this judgment should have been consistent across reviews.

The application of this criterion to studies of *RR* is strikingly different. Three studies of *RR* met the WWC's criteria for review in 2013 (Pinnell, DeFord, & Lyons, 1988; Pinnell, Lyons, DeFord, Bryk, & Seltzer, 1994; Schwartz, 2005). All included random assignment,

---

[8] The WWC policy and procedures handbook states "if an intervention is always offered in combination with a second intervention, any effects cannot be attributed to either intervention separately. Because it is impossible to separate how much of the observed effect was due to the intervention and how much was due to the confounding factor, these studies cannot meet WWC standards"(WWC, 2013d, p. 18).

[9] The sample included both English Language Learners and English speakers. Close reading of the articles shows the authors directly addressed the issue of any possible confound and, based on their results and other work in the field, discounted the possibility. A possible confound was not mentioned in the WWC's 2006 review.

with treatment students receiving one-on-one tutoring with *RR* and control students having alternative treatments in group settings or, in one case (Schwartz, 2005), no additional intervention. However, none of the studies distinguished the impact of the *RR* curriculum from the impact of individualized tutoring. Thus, it is impossible to tell if the *RR* curriculum or extra one-on-one time with an adult produced positive benefits. Even though the studies met the WWC's preference for randomized control trials, they clearly did not meet the criterion of no confounding factors.

Two of the five studies of *RR* accepted by the WWC in 2008 were rejected for the 2013 analysis. One of these works (Iversen & Tunmer 1993) used a design that countered the potential confound of tutoring and curriculum, providing a more appropriate test of *RR*'s efficacy. The authors matched triplets of first grade students on pretest scores and randomly assigned students to 1) the standard *RR* program, 2) a "modified" *RR* program that included explicit instruction in phonological skills, and 3) instruction in the regular classroom program. The major variable of interest was how long children took to reach a specified competency level, the primary goal of *RR*. Students in both the modified and the unmodified program eventually reached this level, but students in the modified program did so more quickly. The standard *RR* program (group 1) was found to be 37 percent less efficient than the modified program (group 2), but more effective than no tutoring at all (group 3). The 2008 WWC report compared the standard *RR* program to instruction in the regular classroom (groups 1 and 3), but did not compare the two interventions that had only one-on-one instruction (groups 1 and 2). In other words, the WWC ignored the analysis that removed the confounding influence of instructional group size, a point we return to below.[10]

Equal Groups at Pretest

The second criterion we examine involves the requirement that the comparison groups be equal at the start of the study.[11] The equivalence of comparison groups is, of course, a key concern of much of the literature regarding experimental design and social scientists have developed sophisticated ways of handling the issue. The WWC does not appear, however, to accept the logic and techniques used throughout the social sciences, ones that are particularly important for use in larger field-based studies.

For instance, the 2008 WWC report on *RM* for beginning reading rejected two studies, both of which found significant positive effects for the DI program, because,

---

[10] The WWC justified its decision to ignore the results when phonics instruction was added to the curriculum "because it was a modified version of the standard program" (Stockard, 2008, pp. 13-14). The opposite reasoning was used in a decision regarding studies that modified the Direct Instruction program, *Corrective Reading.* The WWC accepted a study by Torgesen and associates that used elements of the program even though the authors explicitly stated that "results from this study do not provide complete evaluations of these interventions" (Torgesen, Myers, Schirm, Stuart, Vartivarian, Mansfield, et al., 2006, p. ix; see Stockard, 2013a for more details.).

[11] The WWC handbook states, "at the time the sample is identified (and before the intervention), the groups should be similar, on average, on both observable and unobservable characteristics. This design allows any subsequent (i.e., post intervention) differences in outcomes between the intervention and comparison groups to be attributed solely to the intervention" (WWC, 2013d, p. 9)

although they used a "quasi-experimental design" with both an intervention and a control group, they did "not establish that the comparison group was comparable to the treatment group prior to the start of the intervention" (WWC, 2008a, pp. 2,4). Yet, examination of the studies shows the contrary. One (Brent et al, 1986) adjusted the results, using standard multivariate procedures, for any possible pretest differences between the treatment and control conditions. In another study (O'Brien & Ware 2002), the pretest scores of those receiving *RM* were slightly lower than the scores of students in the control group. This situation is routinely considered a conservative test, for it biases results against the treatment, and is not used as a reason to automatically discount the results.

Another example regarding errors related to this criterion involves Iversen and Tunmer's (1993) study of *RR*. As noted above, the elements of this study that included the confounding factor were included in the 2008 review, but the study was excluded in 2013. An endnote to the 2013 review explains that the change in classification resulted from new standards that require differences between intervention groups at baseline to be smaller than 25% of the pooled standard deviation. However, our calculations, using data from p. 119 of Iversen and Tunmer (1997) indicate that the average difference in pretest scores across the 30 measures reported was only .16 of a standard deviation. The average difference was even smaller (.13 sd) for the two groups that received one-on-one tutoring (the comparison that did not have a confound and that produced negative results for *RR*).

Summary

In each of the examples above the WWC's judgment differs from what the correct application of their own criteria would indicate as well as from what a logical, scholarly analysis would conclude. All of the decisions resulted in a) excluding high quality studies that verified the efficacy of DI programs or the lack of efficacy of *RR* and b) including studies that had faulty designs and promoted *RR*. The probability that such a result would occur by chance is very small.[12]

Interpreting Studies Chosen for Review

The final step in the WWC review process is interpretation of studies and determination of effects. Given how few studies pass the screening criteria for ultimate review it is crucial that these judgments be accurate. Again, however, there are serious errors, and we focus on four examples.

An article by Herrera and associates (1997) was accepted for the WWC review of the use of *RM* for students with LD in 2012. The study compared students who received *RM* to students who received *RM* and additional reading instruction from their *RM* teachers using phonics based movement activities. Not unexpectedly, the students with additional instructional time had higher achievement scores. In the 2012 report the WWC used this finding to conclude that *RM* had negative effects on students' achievement. Clearly, a more reasonable conclusion would be that students with extra instructional time had higher

[12] When data are placed in a four-fold table crossing the scholars' views and the WWC views, the probability associated with a Fisher's Exact Test is .001.

achievement. When the 2012 report was released we objected to this interpretation and requested a quality review. The 2013 report on the use of *RM* with students with LD did not accept the Herrera, et al. article for analysis, stating that it did not provide "enough information about its design to assess whether it meets standards" (WWC, 2013a, p. 7). No mention is made in the 2013 report of its earlier inclusion or why the WWC had altered its view of its acceptability.

Another example involves an article by Cooke and associates (2004) that was found by the WWC to meet evidence standards "without reservations" in both the 2012 and 2013 reviews of the DI program *Reading Mastery* for students with LD. However, in both reports, the WWC analysis directly contradicts the authors' conclusions. Cooke, et al. compared the achievement of 30 students, half randomly assigned to use *RM* and half to use *Horizons,* a modified version of *RM*. The article describes 17 common features 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. The authors found that students in both programs had similar achievement gains over time and report extensive data showing that these gains were significantly greater than those in state and national samples. They concluded that that the slight modifications in *Horizons* had not altered the effectiveness of *RM* documented by other authors. The WWC ignored the comparison to national norms and focused on the lack of differences between the two programs, concluding that that there was no evidence that *RM* was effective. The fact that the students had significantly greater gains than the national norms was not mentioned.[13]

While the faulty interpretation of these studies resulted in non-positive ratings for a DI program, the faulty interpretation of two studies of *RR* contributed to a positive rating in 2008. As noted above, Iversen and Tunmer's (1993) study is the only one of the five accepted by the WWC in 2008 that included controls to address the confounding influence of size of instructional group and curriculum. However, the WWC report only focused on the comparison that retained the confounding elements. Despite the negative conclusions regarding *RR* presented in the article, the WWC reported only a positive effect of the traditional *RR* program.[14]

The second article on *RR* included in 2008 but not in 2013 (Baenen, Bernhole, Dulaney, & Banks, 1997) randomly assigned students to treatment and looked at both short and long-term progress.[15] They reported remarkably little success, especially in the long-

[13] After release of the 2012 report we explained the serious errors of this interpretation of the Cooke, et al. study to the WWC, but it chose to retain its interpretation.

[14] In communication with us the WWC noted that information regarding the more positive results with the addition of phonics was noted in an appendix. The chance of a parent or teacher finding such a detail in an appendix is, of course, slim (Stockard, 2008).

[15] In the 2008 report the Baenen, et al. article was described as a randomized control trial (WWC, 2008b, p. 3), but in 2013 it was described as a "quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent," with an endnote explaining that pretest differences surpassed a newly established criterion of .25 of the pooled standard deviation (WWC, 2013c, pp. 7, 34). Our calculations (using data from p. 170 of Baenen, et al.) indicate that, on average, differences this large only

term, for students in *RR*. Although about one-half of the students in the *RR* tutoring program had successfully reached first grade reading levels at the end of the year, success rates declined in subsequent years. By third grade, there was no difference in achievement scores, needs for retention, special education, or Chapter 1 assignments of *RR* students and other students. The authors concluded that *RR* was very expensive to implement in relation to the benefits it provided. In 2008 we brought this conclusion to the attention of the WWC. The WWC acknowledged the authors' interpretations, but defended their conclusion by stating that their reports "prioritized one-year results" and that the findings regarding the results in later grades were included in a technical appendix (Stockard, 2008, p. 13). Of course, as the research literature firmly indicates, reading skills at the end of third grade are much more crucial in determining eventual success than skills at lower grades.

Summary

The errors at the third stage of the WWC process parallel those at the earlier stages. The interpretations clearly misrepresented the findings of the studies, and resulted in erroneous non-positive reviews to *RM* and erroneous positive reviews to *RR*.

Discussion

The Government Accountability Office has reported that the WWC rejects about 90 percent of the studies that it identifies for possible inclusion (GAO, 2010, p. 13).[16] Given that the WWC's judgments are based on such a small segment of their identified literature, it is crucial that its determinations be as accurate as possible. Unfortunately, our analysis of two curricula indicates that the WWC's conclusions can be deeply flawed. Most disturbing, the misleading reports promote an ineffective program and denigrate one found to be highly effective. Statistical analyses indicate that the probability of these errors occurring by chance is very remote.

The errors that cause these flawed conclusions resulted from problems at each stage of the review process. First, development of lists of studies to examine was incomplete for both curricula, but corrections appear to have been more readily made for the analyses of *RR* than for those of DI. Dozens of studies of DI that could have potentially passed the screening process were not considered and, when provided with suggested citations, the WWC ignored over 80 percent of the documented sources. Second, the criteria that the WWC uses to select studies for full review have been applied in very different ways to the programs. In all cases these discrepancies resulted in the exclusion of high quality studies that support DI programs and the inclusion of flawed studies that support *RR*. Third, there are serious errors in the interpretations of studies selected for review. In all cases, the errors resulted in positive ratings for *RR*, the program deemed by scholars to be ineffective, and non-positive results for DI, the program that scholars have found to be effective. When

---

occurred for one of the cohorts in the analysis and that the authors reported results separately for each cohort. Similar negative results regarding *RR* appeared for each cohort.

[16] The GAO report includes summaries of interviews with researchers about the WWC. Their concerns appear to mirror at least some of those reported in this article.

errors have been brought to the attention of the WWC, the input has usually been dismissed. When the WWC has made changes, these changes have not been acknowledged or explained.

The errors we document could result from a variety of sources, such as poor training or inadequate oversight. Another possibility, however, deserves further attention. Our analysis focused on curricula that represent sharply differing approaches to early reading instruction. It is reasonable to ask if the systematic pattern of mistakes we have documented was influenced by these controversies. Certainly the conclusions presented by the WWC with regard to these two curricula have failed to provide "accurate information on education research" (WWC, 2013a), systematically promoting ineffective curricula and denigrating effective approaches. It is students and the society as a whole that are the true losers in this process, and we urge the educational research community to push for changes that would correct the errors and embody more transparent and appropriate procedures.[17]

[17] We have provided extensive suggestions to the WWC of appropriate policies, focusing on those that would promote accurate assessments of the literature and transparent and accurate review processes. Space limitations preclude including them in this paper, but see Stockard and Wood (2013).

References

Baenen, N., Bernhole, A., Dulaney, C., & Banks, K. (1997). Reading Recovery: Long-term progress after three cohorts. *Journal of Education for Students Placed at Risk, 2*(2), 161-181.

Baker, S., Berninger, V.W., Bruck, M., et al. (2002). *Evidence-based research on Reading Recovery*. http://www.nrrf.org/rrletter_5-02.pdf, retrieved July 19, 2013.

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.

Brent, G., DiObilda, N., & Gavin, F. (1986). Camden Direct Instruction project, 1984-1985. *Urban Education*, 21(2), 138-148.

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.

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.

Government Accountability Office (2010). *Department of Education: Improved dissemination and timely product release would enhance the usefulness of the What Works Clearinghouse* (GAO-10-644). Washington, D.C. GAO.

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.

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.

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

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.

Iversen, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology, 85*(1), 112–126. The

O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth Independent School District. *Journal of Education for Students Placed At Risk*, 7(2), 167-195.

Pinnell, G. S., DeFord, D. E., & Lyons, C. A. (1988). *Reading Recovery: Early intervention for at-risk first graders* (Educational Research Service Monograph). Arlington, VA: Educational Research Service.

Pinnell, G. S., Lyons, C. A., DeFord, D. E., Bryk, A. S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly, 29*(1), 8–39.

Reynolds, M., & Wheldhall, K. (2007). Reading Recovery 20 years down the track: Looking forward, looking back. *International Journal of Disability, Development, and Education, 54,* 199-223.

Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the *Reading Recovery* early intervention. *Journal of Educational Psychology, 97*(2), 257–267.

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. (2013a) Examining the What Works Clearinghouse and its reviews of Direct Instruction programs. Eugene, OR: National Institute for Direct Instruction, Technical Report 2013-1.

Stockard, J. (2013b). 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., & 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.

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.* Retrieved from Institute of Education Sciences, U.S. Department of Education Web site: http://www.ed.gov/rschstat/eval/disadv/title1interimreport/index.html, retrieved September 9, 2013.

Tunmer, W.E., Chapman, J.W., Greaney, K.T., Prochnow, J.E., & Arrow, A.W. (2013). *Why the New Zealand National Literacy Strategy has failed and what can be done about it: Evidence from the Progress in International Reading Literacy Study (PIRLS) 2011 and Reading Recovery Monitoring Reports*. Palmerston North, NZ: Massey University Institute of Education. Retrieved August 29th 2013, from http://www.massey.ac.nz/massey/fms/Massey%20News/2013/8/docs/Report-National-Literacy-Strategy-2013.pdf

What Works Clearinghouse (2008a). *WWC intervention report, Reading Mastery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved August 29th 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/WWC_ReadingMastery_081208.pdf

What Works Clearinghouse (2008b). *WWC intervention report, Reading Recovery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved August 29th 2013, from http://readingrecovery.org/images/pdfs/Reading_Recovery/Research_and_Evaluation/wwc_reading_recovery_report_08.pdf

What Works Clearinghouse (2012). *WWC intervention report, Reading Mastery and students with learning disabilities*. Washington, D.C.: Institute of Education Sciences. Retrieved August 29th 2013, from http://files.eric.ed.gov/fulltext/ED533607.pdf

What Works Clearinghouse (2013a). About us. Washington, D.C.: Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/aboutus.aspx, retrieved September 9, 2013.

What Works Clearinghouse (2013b). *WWC intervention report, Reading Mastery and students with learning disabilities*. Washington, D.C.: Institute of Education Sciences. Retrieved August 29th 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_readingmastery_071712.pdf

What Works Clearinghouse (2013c). *WWC intervention report, Reading Recovery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved August 29th 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_readrecovery_071613.pdf

What Works Clearinghouse (2013d). WWC procedures and standards handbook (Version 3.0). Washington, D.C.: Institute of Education Sciences. Retrieved June 28, 2013, from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

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.

Table 1

WWC Reviews of DI Curricula

Study #	Report Date	Topic/ Protocol	Program	Studies Reviewed	Studies that Fully Met WWC Standards	Studies that Met WWC Standards with reservations	Outcome
1	Sept., 2006	English Language Learners	RM	1	Gunn et al, 2000	0	Potentially positive effects on reading achievement, no information on mathematics achievement or English language development
2	May, 2007	Early Childhood Education	DI	6	0	Cole, et al, 1993	No discernable effects in mathematics, oral language, cognition or print knowledge
3	July, 2007	Beginning Reading	CR	25	Torgesen, et al, 2006	0	Potentially positive effects on fluency and alphabetics; no discernable effects on comprehension; no information on general reading achievement
4	July, 2007	Adolescent Literacy	CR	129	Torgesen, et al, 2006	0	No discernable effects on alphabetics, fluency, or comprehenshion; no information on general literacy achievement
5	Aug., 2008	Beginning Reading	RM	61	0	0	No studies met standards
6	Aug., 2010	Adolescent Literacy	RM	175	Stockard, 2010	Yu & Rachor, 2000	Potentially positive effects on fluency, no discernable effects on comprehension, no information on alphabetics or general literacy achievement

7	July, 2012	Special Needs/ LD	RM	17	Cooke, et al, 2004; Herrera, et al., 1997	0	No discernable effects on comprehension, potentially negative effects on alphabetics, fluency, and writing
8	July, 2013*	Students with Learning Disabilities	RM	22	Cooke, et al., 2004	0	No discernable effects on alphabetics or reading comprehension

*This report is dated 2012 on the WWC website, but is actually a revision of the report that was first issued in July 2012 as described in the text.