The WWC Review Process: An Analysis of Errors in Two Recent Reports

Technical Report 2013-4

Jean Stockard and Timothy W. Wood
Office of Research and Evaluation
NATIONAL INSTITUTE FOR DIRECT INSTRUCTION
JULY 28, 2013

The What Works Clearinghouse Review Process: An Analysis of Errors in Two Recent Reports

The What Works Clearinghouse (WWC) is a federally funded program established in 2002 that evaluates educational interventions and publishes reports and summary ratings. The reports have received extensive criticism, including concerns such as examining only a small proportion of the available evidence, errors in the review process, and a lack of peer review and comparisons of results to related literature. Two WWC reports issued in July 2013 illustrate the severe problems that can permeate the process and result in the dissemination of erroneous conclusions. In one case, the WWC's errors resulted in a positive rating for a program that has been determined, by more inclusive and careful reviews, to be ineffective and inefficient. In the other case the WWC's errors resulted in a negative conclusion regarding a program that has been judged, by more inclusive and careful reviews, to be highly effective. In other words, the errors in the recent WWC reports result in ratings that promote a program found in the established literature to be ineffective and denigrate a program found in all other reviews to be highly effective. These errors illuminate an enormous waste of the nation's resources. But, the true losers are the nation's children, as their schools and educational policy makers are deprived of accurate information on which they can make decisions.

This document describes these errors and preliminary steps to prevent their reoccurrence. The first section reviews the WWC's recent report on the use of the Direct Instruction (DI) program, *Reading Mastery* with students with learning disabilities, which concluded that the program "had no discernible effects" on any of the areas examined. This conclusion is in sharp contrast to all the literature reviews and meta-analyses in the literature. John Hattie's quantitative summary of meta-analyses of Direct Instruction is typical of these results. He summarized the results of four meta-analyses that included DI, incorporating 304 studies, 597 effects and over 42,000 students. He found that the average effect size associated with DI was .59 and noted that the positive results were "similar for regular (d=.99) and special education and lower ability students (d=0.86) [such as those that would be classified as having learning disabilities], … [and] similar for the more low-level word-attack (d=.64) and also for high-level comprehension (d=.54)" (Hattie, 2009, pp. 206-207).[1] No other curricular program reviewed by Hattie showed such consistently strong effects with students of different ability levels, of different ages, and with different subject matters.

The second section focusses on the WWC's recent report on the use of *Reading Recovery* with beginning readers. While the WWC concluded that the program had "positive effects on general reading achievement," the research community has reached an opposite conclusion. In an open letter a group of 31 highly regarded reading researchers summarized

[1] An effect size of .25 has traditionally been considered "educationally important."

"peer-reviewed studies and syntheses of research on *Reading Recovery*" and concluded that "there is little evidence to show that *Reading Recovery* has proved successful with the lowest performing students," the group that it targets (Baker, et al., 2002, p. 1).

Clearly, the two recently released WWC reports directly contradict the overwhelming evidence in the scholarly literature. The first two sections of this report describe the errors that led to the WWC's faulty conclusions, and the third section discusses the implications of the analyses, with special attention to the WWC's published policies and recommendations for action.

WWC Reviews of Reading Mastery for Students with Learning Disabilities

In July 2012, the WWC issued a report regarding the use of *Reading Mastery (RM)* for students with learning disabilities, using the results of only two studies with a total of 113 students. In contrast to the conclusions of all of the meta-analytic studies and comprehensive literature reviews in the area, the WWC concluded that *RM* had potentially negative effects. NIFDI's office of Research and Evaluation found very serious errors in the analysis of these two studies and conveyed these concerns to the WWC. The WWC subsequently removed the report from its website.[2] A second WWC report on the use of *Reading Mastery* for students with learning disabilities was issued in July 2013. Surprisingly, this report is also dated July 2012 and no reference is given to the fact that it is a revision of the report that was posted earlier.[3] The 2013 analysis retains many of the errors that were in the 2012 report and does not incorporate any of the carefully developed analyses and reviews that were provided to the WWC by NIFDI.

As described in much greater detail in other writings (Stockard and Wood, 2012; Stockard, 2013a), the 2012 WWC report looked at only a fraction of the studies that have examined the use of Reading Mastery with students with reading difficulties, including learning disabilities. This problem was not corrected in the 2013 report. Table 1 summarizes the number of studies included in the two analyses. The 2012 report by the WWC examined 17 articles and determined that only 2 met their criteria for inclusion. The 2013 report examined 22 articles and determined that only 1 met their criteria. Two of the studies added in 2013 were meta-analyses, and three were research reports.

An article by Herrera and associates (1997) was accepted for review in 2012, but was described in 2013 as not providing "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 the change of placement. The study compared students who received only *RM* to students who received *RM* and additional reading instruction from their *RM* teacher using a system of 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'

[2] The NIFDI research staff assumes that the removal was in response to the concerns that they had submitted, but the research office was never notified that the removal had occurred or told of the reasons for the decision.
[3] This document continues to refer to the reports by the date at which they were issued. A copy of the 2012 report is still posted at ERIC and, of course, the original report was publicized and noted in blogs and various media outlets.

achievement. A more reasonable conclusion, of course, would have been that students with extra instructional time had higher achievement; and that is why NIFDI objected to its inclusion. It is impossible to tell from the 2013 report what judgment the WWC made of the study in its revised analysis.

Table One

Disposition of Studies Reviewed by the WWC for the Analyses of Reading Mastery with Students with Learning Disabilities

	2012	2013
Met without reservations	1	1
Met with reservations	1	0
Failed to meet evidence standards	1	1
Not an efficacy study	6	8
Fewer than 50% of students with LD	4	6
Design not within protocol	4	5
Insufficient information	0	1
Total	17	22

An article by Cooke and associates (2004) was found by the WWC to meet their evidence standards "without reservations" in both 2012 and 2013. However, in both reports, the analysis of the results is deeply flawed and directly contradicts the authors' conclusions. Cooke, et al. compared the achievement of 30 students using *RM* with those who used *Horizons*, a slight modification of the *RM* program developed by the author of *RM* and his associates. They found that students in both *RM* and *Horizons* had similar achievement gains over time and that these gains were significantly greater than those in state and national samples. In other words, they concluded that both of the programs were effective and 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 instead focused on the lack of differences between the two programs. They concluded, in both the 2012 and the 2013 analyses that the lack of difference in results between *RM* and its modified version, *Horizons*, indicated that there was no evidence that *RM* was effective. The summary judgment, included in the body of the report is "When compared to another Direct Instruction intervention, *Horizons*, *Reading Mastery* was found to have no discernible effects on alphabetics and reading comprehension for students with learning disabilities." The fact that the students had significantly greater gains than the national norms is not mentioned at any point.

As noted above, the NIFDI research office has completed two extensive reviews of studies that should have influenced the reanalysis of the WWC's 2012 report (Stockard and Wood, 2012; Stockard, 2013a). The NIFDI analyses were prompted by concerns over the sharp differences between the WWC's conclusions and the extant scholarly literature, as noted in the introduction. The 2012 analysis (Stockard and Wood) includes a lengthy bibliography of works that should have been included in the WWC review. The more recent report (Stockard, 2013a) documents specific errors in a substantial proportion of the WWC's decisions

regarding both inclusion and exclusion of studies from review in all of their reports on DI materials. It also includes a meta-analysis of the results of 21 research studies on the use of *Reading Mastery* with students with reading difficulties. Details regarding the studies' designs and conclusions are given, and effect sizes associated with the results are calculated and analyzed.[4] The report was given to the WWC office; and the coordinator in the WWC Learning Disabilities Review acknowledged its receipt in February, 2013, noting that "it was very helpful in our review of *Reading Mastery*."

Surprisingly, however, the July 2013 analysis of *RM* failed to review the vast majority of the studies that the NIFDI reviews found to be appropriate. For instance, six of the literature reviews and meta-analyses listed in Stockard and Wood (2012) were not in the 2012 WWC report. One could expect that these six documents would be consulted by the WWC in a revision. However, the WWC lists only two of these six reviews as consulted in 2013. Three of the 21 research articles in NIFDI's 2013 meta-analysis were included in the WWC 2012 listing (albeit with inappropriate interpretations by the WWC). Of the remaining 18 studies in the NIFDI meta-analysis only two were added to the 2013 review. One of the articles added was the only one in the list that had been authored by the NIFDI research office (Stockard, 2008), and the other 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 appears to have ignored well over half of the material that directly addressed the use of *Reading Mastery* with students with learning disabilities, even when presented with extensive analyses that demonstrated its relevance. Moreover, their decisions about which studies to include appear to be highly selective.

The WWC's omission of two large, well designed, federally funded studies that were included in NIFDI's listings is especially troubling: the work of Gunn and associates (2000, 2002, 2005) and a study by Kamps and associates (2003). The Gunn, et al. work included random assignment of students to treatment, a large number of assessments, and follow-up of students for several years. The Kamps, et al work used sophisticated statistical analyses to examine growth in learning over time in a variety of schools. (See Stockard, 2013a, pp. 18-21 for more details on these studies.)[5] None of these works were included in the list of studies reviewed by the WWC in either 2012 or 2013, even though they figured prominently in the NIFDI analyses.[6]

The errors in the WWC process in the reviews of *Reading Mastery* have resulted in negative ratings being given to a program that has been found in all other expert reviews to be highly

[4] Mixed models were used to adjust for multiple effects within studies and varying sample size. The resulting average effect size was .37, well above the criterion of .25 commonly used to denote educational importance, but somewhat lower than in other meta-analyses in the DI literature.

[5] The WWC may object to including these studies because the term "learning disabled" is not explicitly mentioned. However, as pointed out in the NIFDI reports, the definition of learning disabled varies over both time and locale. It is arguably much more appropriate to look at all literature regarding general issues of reading difficulties or "struggling readers" rather than a varied and imprecise concept such as "learning disabled."

[6] It is, of course, possible, if not probable, that the WWC would decide that these studies do not meet their protocols for review. As described more fully in Stockard (2013), the WWC criteria for selection of studies appear to be biased against the inclusion of larger, field-based studies such as these. It is beyond ironic that studies such as these can receive very large amounts of grant funding from the federal government yet fail to pass the screening criteria of the WWC.

effective (e.g., Adams & Engelmann, 1995; Borman, Hewes, Overman, & Brown, 2003; Hattie, 2009; White, 1988). In contrast, the errors in the WWC process in the reviews of *Reading Recovery* have resulted in positive ratings being given to a program that expert reviewers have rated as ineffective and inefficient.

WWC Reviews of Reading Recovery for Beginning Reading

In July 2013 the What Works Clearinghouse (WWC) issued a report regarding the impact of the *Reading Recovery (RR)* program for beginning readers. The report is described as an update, although no reference or comparison is made to the original report issued in 2008. As described by the WWC,

> *Reading Recovery*® is a short-term intervention that provides one-on-one tutoring to first-grade students who are struggling in reading and writing. The supplementary program aims to promote literacy skills and foster the development of reading and writing strategies by tailoring individualized lessons to each student. Tutoring is delivered by trained *Reading Recovery*® teachers in daily 30 minute pull-out sessions over the course of 12–20 weeks. (WWC, 2013b, p. 1)

The 2008 analysis identified approximately 100 studies that investigated the effects of *Reading Recovery* on the reading skills of beginning readers, while the 2013 report identified about twice that many. None of the additional studies were determined to meet evidence standards with or without reservations. The WWC reported that five studies of *RR* met their criteria for review in 2008, but that only three of these studies met the criteria in 2013. As with the 2012 and 2013 analyses of *Reading Mastery*, the WWC makes no explicit comparison between the 2008 and 2013 *RR* analyses and does not explain why two studies were reclassified in the later publication.

Examination of the five studies that were accepted for review reveals that none of them, as interpreted by the WWC, provides an adequate test of the program's efficacy. Specifically, none of the studies distinguished the impact of the *RR* curriculum from the impact of having individualized tutoring. In other words, none of these studies controlled for the so-called "Hawthorne Effect," in which simply the extra attention obtained from an experimental condition can produce change. [7] Thus, it is impossible to tell from these studies if the *RR* curriculum or simply extra one-on-one time with an adult produced positive benefits. The WWC makes no mention of this confounding effect in their analysis. The designs of the three studies in the 2013 review are briefly described below. This description is followed by a discussion of two studies that were included in the 2008, but not in the 2013, review and then a section that summarizes the patterns of errors.

[7] As described more fully below, one of the articles accepted for the 2008 review did include elements of the design that controlled for this effect, but that part of the study (which reflected negatively on RR) was ignored by the WWC in its review.

Studies Included in the 2013 WWC Review of Reading Recovery

Brief descriptions of the three studies included in the 2013 review of *Reading Recovery* are given below along with an analysis of serious issues with the WWC's interpretations of the findings. These analyses are based on the descriptions given in the WWC report and examination of the original articles.

1) 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.[8]

Low achieving students in first grade classes were randomly assigned to participate in *RR* or to the comparison group. All students attended regular education classes. The *RR* students had an additional 30 minutes of individualized instruction with an *RR* teacher, while those in the comparison group had additional instruction in "an alternative compensatory program" described as having "minimal individual-level instruction." In other words, the differences between the groups involved two elements: 1) one-on-one tutoring versus small group instruction and 2) the *RR* curriculum versus other types of compensatory material. There is no way within the design to separate these effects; it is impossible to distinguish which of these variables produced differences between the groups. In other words, the impacts of the two variables are confounded. While the WWC concluded that the results indicated a positive effect of the program, it is just as likely that the result came from the individualized attention, a classic Hawthorne effect.

2) 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.

In this study low achieving first grade students from four schools were randomly assigned to receive *RR* individualized tutoring or to have the reading intervention used at their school. The interventions for the comparison group were described as their "regular reading program" and supplemental services including "teachers reading aloud as well as group reading" (WWC, 2013b, p. 26). The comparison students were taught in group settings, while the intervention students had one-on-one tutoring. Like the 1988 study described above, this design confounds the size of group and the curriculum and there is no way in which one can determine which factor produced the results.

3) 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.

In this study low achieving first grade students were randomly assigned to receive Reading Recovery or to be in a wait-listed control group (receiving the program in the second half of the school year). The intervention group had their regular reading program plus the half hour daily intervention of Reading Recovery, while the comparison group "received instruction in their regular classroom but no additional supplemental services" (WWC, 2013b, p. 27).

[8] The authors were unable to find this monograph and, instead, consulted Pinnell (1989), one of the other sources listed in the WWC report.

Thus, the intervention group had extra instruction in a tutoring setting, while the comparison group had no extra instructional time nor individualized instruction. Again, the WWC's conclusion of positive results for the *RR* curriculum is faulty for it is impossible to determine if the effect is due to the one-on-one attention, additional study time, or the *RR* curriculum.

The 2008 WWC Review of Reading Recovery

As noted above, the 2013 WWC report on *RR* was an updated version of a report issued in 2008. The 2008 version included all of the studies noted above plus two other reports. An article by Baenen and associates (1997) was found to meet all of the standards for inclusion, and one by Iversen and Tunmer (1993) met the standards "with reservations." In the 2013 report both of these articles are listed, but they were found not to meet the standards because they used "a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent" (WWC, 2013, pp. 7 and 10). No mention was made of their acceptance in the previous analysis or of the reason for the change in their classification.

While the 2008 version of the *RR* report cited both articles as showing a positive impact of the program, the actual content of the articles provided contrary conclusions. For instance Baenen and associates (1997) reported positive short-term results with the program at the end of first grade, but that the positive impacts disappeared in later years. They concluded that the cost of the program was very expensive relative to any benefits it provided. In correspondence regarding this issue the WWC acknowledged the negative conclusions of the study's authors but defended their positive assessment 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 (See Stockard, 2008, pp. 13-14).

The article by Iversen and Tunmer (1993) is the only one of the five accepted by the WWC that dealt effectively with the potential confound of size of the instructional group and curriculum. The authors matched triplets of first grade students on pretest scores and randomly assigned students to one of three groups: 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 to Iversen and Tunmer was how long children took to reach a level of competency where they could discontinue special tutoring, the major goal of a tutoring program such as *RR*. Students in both the unmodified *Reading Recovery* program and the modified program (including instruction in phonologically based elements) eventually caught up with the other children, but the students in the modified program were able to discontinue tutoring much earlier. The standard *Reading Recovery* 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 report of the WWC only focused on the comparison between group 1 (students in the regular *RR* program) and group 3 (those with no tutoring and their regular instruction). In other words, the WWC ignored the analysis that removed the confounding influence of instructional group size. The 2008 WWC report indicated that Iversen and Tunmer's study showed a positive effect of *RR*, while their actual conclusion was the opposite. When questioned about this analysis, 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." Again, they noted that the information regarding the more positive results with the addition of phonics were noted in an appendix. (See Stockard, 2008 for a more complete discussion and documentation.)

Summary

Table Two summarizes the patterns described in the previous two sections. The WWC concluded that all five of the studies demonstrated positive impacts of the *RR* program. However, in all cases, their conclusion ignored the confounding influence of the size of instructional group and curriculum. Moreover in two of the analyses their conclusions were contrary to those reached by the authors, who reported negative results. In short, the WWC's conclusions about the positive effects of *RR* are not based on solid evidence. Four of the five studies do not separate the influence of size of instructional group and curriculum. The only study that controlled for the confounding influence of size of instructional group and curricular approach was Iversen and Tunmer (1997). It found negative results for *RR*, but those negative results were ignored in the 2008 WWC report and the study was omitted from review in the 2013 report.

Table Two
Summary of WWC Reviews of Reading Recovery Studies Accepted for Review

			Conclusions Regarding *RR* Effects	
Study	Year of WWC Review	Design Confound?	By Authors	By the WWC
Pinnell, et al, 1988	2008, 2013	Yes	Positive	Positive
Pinnell, et al, 1994	2008, 2013	Yes	Positive	Positive
Schwartz, 2005	2008, 2013	Yes	Positive	Positive
Baenen, et al, 1997	2008	Yes	Negative	Positive
Iversen & Tunmer, 1993	2008	Partly*	Negative	Positive

*The WWC reached its positive conclusion by ignoring the results from the element of the Iversen and Tunmer study that did not involve the confounding influence of size of instructional group and curriculum. In other words, their positive rating was based on the parts of the study that involved the confounding influences.

Note that if the results with the flawed designs were omitted, the WWC's conclusions would alter quite dramatically. Only one of the five articles reviewed over the two reports would be included in the analysis, and within that report only the elements that allowed analyses that removed the confounding effects of size of the instructional group and curriculum would be examined. The appropriate conclusion of such a summary would be that *Reading Recovery* is not as effective as tutoring programs that incorporate phonics. Such a conclusion would

be in line with that reached by the large group of international reading experts quoted above (Baker, Berninger, Bruck, et al., 2002).

Implications and Discussion

Many of the problematic issues described in the analysis of the work with *Reading Mastery* are also evident in the updated report on *Reading Recovery*. In both cases only a fraction of the studies that were initially examined were retained for review. The WWC's conclusions about studies they did select for review are misleading and counter the conclusions of experts in the field and well regarded literature summaries and meta-analyses. Misinterpretations of the actual conclusions of the studies are even more alarming, both because they have been documented on multiple occasions in reports submitted to the WWC and because they directly contradict the conclusions of independent experts. In this section we first discuss problems with the *RR* and *RM* reports in light of the stated procedures of the WWC and then discuss recommendations for future action and change.

WWC Policies in Practice

According to the WWC's 2013 Procedures and Standards Handbook "It is critical that educators have access to the best evidence about the effectiveness of education programs, policies, and practices in order to make sound decisions" (WWC, 2013c, p. vi). Unfortunately, as described above, at least some of their reports provide incomplete and misleading analyses and thus do not provide this "best evidence." Four general problems with the WWC's reviews analyzed in this report can be highlighted: 1) issues regarding the selection of studies for review, 2) errors and misinterpretations of studies that are reviewed, 3) a lack of transparency in procedures and the review process, and 4) an apparent lack of competent quality control and expert review. Each of these problems has been discussed to at least some degree in other writings by the NIFDI office, and the discussion below primarily focuses on how they apply to the reviews of *RM* for students with LD and *RR* discussed above.

Incomplete and Biased Selection of Studies for Review – The WWC touts its "comprehensive coverage of the relevant literature" (2013c, p. vi), yet the WWC's selection criteria appear to greatly limit their analysis of relevant and critical studies. One issue in this area involves inadequate and incomplete searches for relevant articles. As described above, both the 2012 and 2013 studies of *RM* looked at only a fraction of the available studies, ignoring many that were on the lists of studies on the effect of *Reading Mastery* on students with learning difficulties (Stockard and Wood, 2012) and a detailed analysis of these studies (Stockard, 2013a), both of which were submitted to the WWC in response to the original report. The absence of these additional studies from the updated report appears to indicate a problem in the WWC's current practices and a departure from the stated procedures. The current 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 (4). WWC procedures dictate "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" (7). If these procedures were accurately followed the WWC should have been aware of several more relevant studies regarding *RM*, even without the assistance of NIFDI submitting their lists of relevant works.

Data presented above also indicate that the reviews of *RR* involved substantially more studies than those for RM. Over 200 citations were listed for the *RR* review. Of the approximately 100 studies that were added to the RR review for 2013, almost half were published before 2008, the date of the first review, indicating that their initial search of the literature for RR was also incomplete. In contrast, fewer than two dozen were listed for the *RM* review, even though the WWC acknowledged receipt of lists of additional studies as noted directly above. The authors do not know the extent of the *RR* literature, but it is not unreasonable to hypothesize that a higher proportion of the extant *RR* literature than the *RM* literature was examined.

In addition to incomplete searching of the literature, the WWC criteria for acceptance of studies should be questioned. As explained above, very small proportions of the literature found in the reviews are actually accepted for analysis. Of the 17 studies in the 2012 study of *RM* only two were accepted and, of the *22* studies in the 2013 report, only one was accepted. The authors are far less familiar with the literature regarding *Reading Recovery*. However, the fact that only three of over 200 articles considered were found to meet the review criteria suggests that similar issues may be involved with that report. Such strict criteria would be appropriate if they resulted in more accurate examination of the literature and summary conclusions. However, an empirical analysis of this question (Stockard, 2013a) found no indication that the criteria were related to the effect sizes of studies. In addition, the WWC does not appear to provide any research-based justifications for their decision process. In short, the limited and selective nature of the WWC reports seems in direct opposition to their claim of "comprehensive coverage of the relevant literature" (WWC 2013c, p. vi). Both their incomplete searches of the available literature and their highly restrictive screening process limit the use of relevant studies, working directly against their goal of understanding "what works" in education.

Errors in Interpretation and Analysis – The WWC's analyses of reports appear to have numerous errors. Some of the reports have misinterpreted and/or avoided author conclusions about the studies. Two prominent examples are the WWC's interpretations of the Cooke, et al. study of *Reading Mastery* and *Horizons* in the 2012 and 2013 reports and the WWC's interpretation of the Iversen and Tunmer study in the 2008 report on *Reading Recovery*. While the authors of these studies found evidence of positive effects for *RM* and negative effects for *RR*, the WWC's conclusions were just the opposite. Errors in other studies of Direct Instruction programs have been documented in other writings (e.g. Stockard, 2008). The numerous errors are especially troubling given the WWC's depiction of its work as the source of "credible and reliable evidence" ("about us" web page - http://ies.ed.gov/ncee/wwc/aboutus.aspx)

There are also disturbing inconsistencies in the interpretations of studies that were accepted for review, and it appears clear from the above analysis that different criteria have been used in the analyses of Reading Recovery and Reading Mastery. Two examples involve

specific decisions in the two reviews, and the third involves more general issues in the ways in which research designs are evaluated. In all cases, the errors have resulted in misleading and erroneous negative reviews for *Reading Mastery* and misleading and erroneous positive reviews for *Reading Recovery*.

The first example involves the decision to omit the Herrera et al (1997) study from a review of *Reading Mastery*, but a failure to omit studies with equally flawed designs from the reviews of *Reading Recovery*. The decision of the WWC to omit Herrera et al (1997) from review in 2013, as described more fully above, is appropriate. The intervention and comparison groups were not comparable because the intervention group had extra instructional time and it was impossible to determine whether changes resulted from extra instructional time or the intervention. However, all three of the studies accepted for the 2013 *Reading Recovery* report have similar problems and should also have been omitted. As described above, it is impossible to tell from those studies if the results appeared because the intervention students had extra time in one-on-one tutoring sessions or if they appeared because of the nature of the curricular material. Ironically, one study of *Reading Recovery* that avoided this confound (Iversen and Tunmer, 1997) was accepted by the WWC for review in 2008. However, in the 2008 report the WWC ignored the findings from the part of the study that removed the confounding effect. In 2013 the study was totally omitted from review.

The justification given for the WWC decision to disregard the comparison in the Iversen and Tunmer (1993) study of *Reading Recovery* in its 2008 analysis is evidence of another type of inconsistency in the reviews of the two programs. As explained above and examined more fully in Stockard (2008), the WWC justified its failure to compare results of students tutored with *Reading Recovery* and those tutored with the program to which a phonics element had been added "because it [the version with phonics added] was a modified version of the standard program" (Stockard 2008, pp. 13-14). The Cooke et al (2004) article provides extensive details demonstrating the ways in which *Horizons* is a slightly modified version of *Reading Mastery* – in fact with far fewer modifications than were in the phonics-supplemented *RR* program. Yet the WWC chose to treat *Horizons* and *Reading Mastery* as two distinct programs. This discrepancy is even more disturbing given the types of results that were involved. The study comparing *RR* and *RR* plus phonics found that the modified version was significantly more effective – a conclusion in direct contradiction to the finding presented by the WWC. The study comparing *RM* and *Horizons* found that both produced significantly greater gains than occur in national and state samples, but the WWC ignored this positive result and instead stated that the lack of difference between *RM* and its modified version indicated there was no evidence of effectiveness of *RM*. The decision of the WWC is, in both cases, illogical and, more importantly, clearly misrepresents the research findings of both *RR* and *RM*.

The third example involves more general issues in the reasons given for rejecting or accepting studies for review and the WWC's very narrow understanding (or, perhaps, a broader misunderstanding) of basic logical issues in research design. As noted above, there were fatal flaws in the designs of all of the *RR* studies that were reviewed and these flaws were not acknowledged in any way. While the WWC gives strikingly little detail regarding reasons for determining that studies are ineligible for review, at least some of the decisions

are in sharp contrast to the literature. For instance, they seem to reject all studies that use a cohort control group design (e.g. SRA/McGraw Hill, n.d., 2006, 2009, WWC 2013a, p. 8), claiming that they did "not use a comparison group design," a statement that is clearly incorrect. The classic Campbell and Stanley literature on research design and a more recent publication in a leading, high quality social science methodology journal describe in great detail why this type of design is well suited, with both high external and internal validity, for studies in organizational settings such as education (see Stockard, 2013b for more extensive discussion of the general issues related to research design).[9] Again, the end result has been decisions that give erroneous positive rankings to *RR* and erroneous negative rankings to *RM*.

Lack of Transparency and Clear Methods for Corrections of Errors – The WWC claims their "systematic review process is the basis of all WWC products, enabling the WWC to use consistent, objective, and transparent standards and procedures in its reviews, while also ensuring comprehensive coverage of the relevant literature" (WWC, 2013c, p. vi). Clearly, however, this claim of transparency of procedures and reviews must be called into question given the analyses presented above. The reports on *Reading Mastery* and *Reading Recovery* issued in July 2013 are revisions of earlier reports and both involve substantive differences from the first editions. As described above, the two *RM* reports have slightly different conclusions, with the 2012 analysis reporting negative impacts and the 2013 analysis reporting no discernible effects. The 2013 report on *RR* also differs substantially from the 2008 version, with only three of the articles accepted for review retained for the 2013 analysis. Despite the differences between these versions of the reports the WWC does not discuss or, much less, acknowledge the changes. A "consistent, objective, and transparent" system should note all relevant changes between the two reports.

Even more disturbing is the fact that the 2013 updated RM report has been backdated as 2012 and includes no reference to its previous publication. One could suggest that the WWC is trying to portray the 2013 analysis as the original. By not acknowledging the original publication the WWC admits to no errors in their previous report and also appears to market the report as a first edition. The intention of the backdating is, of course, not clear, but its effect of hiding errors in the process is troubling and works directly against a goal of transparency. The possibility of the backdating being used to cover up errors in the past should be of great concern to the WWC, the government agencies sponsoring their work, and the general public. The backdating may have been a simple mistake in the editing process. Yet, with the rigorous standards and review processes that are supposedly employed by the WWC it is hard to imagine this went unnoticed by any of their employees, especially members of their statistical, technical and analysis team, the focus of the quality control and review procedures.

Lack of Quality Control and Skilled Review – The WWC describes their statistical, technical and analysis team as "a group of highly experienced researchers who consider issues requiring higher-level technical skills, including revising existing standards and developing new standards" (A.3). This team is also described as consulted on issues that arise during

[9] The WWC was provided with the Stockard (2013b) article and its use in interpreting the SRA/McGraw Hill studies was demonstrated in the Stockard (2013) meta-analysis. It is unclear why they chose to ignore this information.

the review process, and it is logical to conclude that they are responsible for issues of quality control and expert review. Given the extensive errors noted in the documents reviewed in this report, the effectiveness of this team must be questioned. It is hard to imagine how a competent quality review team could fail to note the limited literature reviews and the errors in interpretations that permeate the reports discussed here. It is also hard to understand how a competent review group would allow, as happened with the conclusion regarding *RM*, a decision to be based on one study of 30 students, when given a meta-analysis of 21 relevant studies involving hundreds of students that reached an opposite conclusion.

Recommendations

The present analysis has focused on only four WWC reports. NIFDI's office of research and evaluation has documented numerous errors in other work by the WWC, including those involving several other reports on Direct Instruction program. Based on these reviews, there is no reason to expect that errors described here are limited to just the four reports examined. Three types of actions appear appropriate and needed.

First, the WWC should immediately withdraw the reviews of *Reading Mastery* and *Reading Recovery* that were posted in July and post announcements on its website of the flawed conclusions. Revision of the *RM* report must take into account the full literature base and present an accurate analysis of the Cooke, et al. study. Revision of the *RR* report must consider the design flaws and confounding effects in the studies that were accepted for 2013 and present an accurate analysis of the Iversen and Tunmer study. Both revised reports should go through extensive review by outside analysts. The NIFDI research office will ask that the reviews be removed from the website and that the WWC instigate immediate Quality Review processes for both reports.

Second, it is important to try to understand why the problems described in this report occurred. Possible reasons could include inadequate academic training of WWC staff, problems with poor management and oversight of the review process, and/or willful misrepresentation of the literature base and study results. From the information currently available it is impossible to tell what has led to the erroneous and misleading conclusions. To help understand the source of the problems and to prevent them from occurring in the future, NIFDI's research office is submitting a Freedom of Information Act request for information on the review process. Transparency of the review process is not just part of the stated policies and procedures of the WWC. It is required of federal agencies. It is hoped that the release of information about the review process will help prevent such serious errors in the future.

Third, it is clear that the current WWC procedures and policies are not sufficient to guarantee accurate analyses. In the spring of 2013 the WWC invited public input on a revision of their policies and procedures. Table Three lists the recommendations that NIFDI submitted to the WWC at that time. The recommendations developed from careful review of the reports on Direct Instruction programs conducted by the WWC as well as a statistical analysis of the impact of the WWC's selection criteria on reported results and comparison of the WWC approach with standard methodologies used in the social science (Stockard and

Wood, 2012; Stockard, 2008, 2013a). The recommendations reflect the issues discussed above and involve encouraging a much larger scope of studies in the review process; ensuring that reviewers are knowledgeable in substantive areas reviewed; extensive checks on results, both within the WWC and by independent peer reviewers; full transparency of the process of review; and comparison of results with previously published literature reviews and meta-analyses.

The WWC states its mission is "to be a central and trusted source of scientific evidence for what works in education" (2013c, p. vi). Unfortunately the practices and conclusions of the WWC do not always follow the expert opinions of educators and researchers who rely on scientific practices and data to determine what works in education. On multiple occasions the WWC has released reports with misleading conclusions and examined only a fraction of the relevant studies. The practices of the WWC must again be called into question following the July 2013, release of the updated reports on *RR* and *RM*. Both reports continue to feature misleading and harmful information on what works in education.

Table Three
Recommendations Given to the WWC by NIFDI as Part of the Public Input Process, April, 2013

- The preference for small, tightly restricted, randomized control trials effectively excludes most field-based studies of larger populations and those that use advanced statistical methods for controls. Such larger, field based trials are especially important for ensuring external validity. The social science community increasingly uses such techniques and approaches, and the WWC restriction greatly limits the literature included in reviews. Instead of simply rating studies' quality by the nature of the research design, elements related to sample size, statistical significance, substantive significance, and length and fidelity of intervention should be included, with global ratings reflecting the preponderance of evidence regarding interventions from all available data.

- The focus on narrow curricular programs (e.g. specific titles of reading series), rather than instructional approaches (e.g. Direct Instruction reading) can misrepresent the nature of curricular approaches and introduce artificial distinctions between elements of a curricular approach. It would often be more appropriate and accurate to use a broader frame in which to capture relevant studies.

- The WWC should ensure that reviewers are knowledgeable in the substantive areas that they are reviewing as well as in methodological details. For instance, Direct Instruction programs embody procedures for reinforcement of behaviors and classroom behavioral management, but at least one review discarded a study because it mentioned these elements of the program, claiming that the behavioral elements were a "confound" to the approach.

- The WWC generally excludes studies that were published more than 20 years ago from review. This decision should be justified with research that clearly demonstrates that the ways in which children learn have altered over time. (I know of no such

research and have looked quite carefully for this evidence.) If there is no such research, then this ban should be altered and all of the literature related to a curriculum should be reviewed.

- The literature searches, at least with respect to Direct Instruction programs, appear to be incomplete. In areas with extensive meta-analytic work and literature reviews, all of the studies cited in those reviews should be examined and full listings of the reviews, meta-analyses, and other means of gathering literature should be reported.

- Decisions regarding inclusion or exclusion of studies from review should be made independently by at least two reviewers. Discrepancies in decisions should be carefully analyzed and resolved by a third party. I have found numerous errors in inclusion and exclusion decisions, suggesting that much better quality control is needed.

- All reports should be subjected to peer review that is independent of the WWC before posting. These peer reviews should be available on the WWC site.

- The conclusions of reports should be compared with the meta-analytic literature and other research syntheses. When the WWC conclusions differ from the established literature, the WWC should take extensive measures to understand the discrepancies, including correspondence with the authors of the meta-analyses and individual studies and reports of the ways in which the conclusions differ. These analyses should be publicly available.

- Before posting, the draft reports should be sent to the authors of the programs and other organizations, such as NIFDI, that are familiar with the programs and research. Input should be sought regarding accuracy of the reports, and this input should be used to revise the reports before posting to the web. When the input is not used in revisions, it should be publicly posted to allow consumers to independently evaluate the reports.

- The decision making process of the WWC should be fully transparent. Quality reviews should include analyses by reviewers who are independent of the WWC. As part of the transparency effort, the WWC should make available the requests for quality review that have occurred, the procedures used to address them, the decisions that were made, and the reasons for these decisions. I believe that federal law requires such transparency.

- A cycle for review and revision of reports should be established and posted. This is especially important when there have been serious criticisms of reports, including requests, either formal or informal, for quality review.

References

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

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.

Baker, S., Berninger, V.W., Bruck, M., et al. (2002). Evidence-Based Research on Reading Recovery http://www.nrrf.org/rrletter_5-02.pdf- downloaded 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.

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.

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.

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.

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

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.

Pinnell, G. S. (1989a). Reading Recovery: Helping at-risk children learn to read. *The Elementary School Journal, 90*, 161–183.

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 2013 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.

What Works Clearinghouse (2008). *WWC Intervention Report, Reading Recovery and Beginning Reading*. Washington, D.C.: Institute of Education Sciences.

What Works Clearinghouse (2013). *WWC Intervention Report, Reading Recovery and Beginning Reading*. Washington, D.C.: Institute of Education Sciences.

What Works Clearinghouse (2012). *WWC Intervention Report, Reading Mastery and Students with Learning Disabilities*. Washington, D.C.: Institute of Education Sciences.

What Works Clearinghouse (2013a). *WWC Intervention Report, Reading Mastery and Students with Learning Disabilities*. Washington, D.C.: Institute of Education Sciences.

What Works Clearinghouse (2013b). *WWC Intervention Report, Reading Recovery and Beginning Reading*. Washington, D.C.: Institute of Education Sciences.

What Works Clearinghouse (2013c). *WWC Procedures and Standards Handbook*. Washington, D.C.: 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.