The What Works Clearinghouse Beginning Reading Reports and Rating of *Reading Mastery*: An Evaluation and Comment

Jean Stockard, Ph.D.
Director of Research
National Institute for Direct Instruction
Eugene, Oregon

September 2008

Technical Report 2008-04

PO Box 11248 • Eugene, OR 97440
1-877-485-1973

The What Works Clearinghouse Beginning Reading Reports and Rating of *Reading Mastery*: An Evaluation and Comment

NIFDI Technical Report 2008-4
September, 2008
Executive Summary

The What Works Clearinghouse is a federally funded program established in 2002 that evaluates educational interventions on the basis of the "rigor of research evidence" and provides summary ratings on its website. Ratings have appeared only since 2007, and some of the work of the Clearinghouse has already been subjected to harsh criticism. This paper adds to the body of critiques.

Direct Instruction's *Reading Mastery* curriculum is a comprehensive core reading program that has been found, in decades of research, to be highly effective. Yet, the WWC's recent review of *Reading Mastery* concluded that no studies met their evidence standards and they were "unable to draw any conclusions based on research" regarding its effectiveness.

A careful examination of the research cited by the WWC found that this conclusion was not well founded. The WWC failed to examine close to 100 research studies that were cited in other well-known reviews of the literature. In addition, reviews of studies that were included were flawed. For instance, one large, well-regarded study was rejected because training of teachers to use the curriculum was considered a "confound" to the use of the curriculum itself. Other studies were dismissed because the treatment (*Reading Mastery*) and control groups were unequal at baseline. Yet, these differences were statistically controlled and, in two cases, the treatment groups had lower pretest scores but equal and higher posttest scores.

Examination of reviews of other curricula indicates that at least some of the material posted on the website is misleading. Summaries do not reflect the conclusions of the articles and, in at least one case, are exactly opposite to the published conclusions.

Attempts have been made to communicate these concerns and recommendations for change to the WWC and Mathematica Policy Research, Inc., which administers the Clearinghouse. These concerns and suggestions, which parallel those of other researchers, have been rejected.

Given the very serious problems in the review procedures and policies, the NIFDI research staff suggests extreme caution in the use of any ratings or conclusions from the WWC.

Table of Contents

The What Works Clearinghouse Beginning Reading Reports and Rating of *Reading Mastery*: An Evaluation and Comment

This document discusses serious concerns with the policies, procedures and judgments of the federally funded What Works Clearinghouse, focusing both on a recent Clearinghouse review of the Direct Instruction program *Reading Mastery* as well as more general issues regarding their procedures and conclusions. The report begins with a brief description of the stated mission of the WWC and criticisms of their procedures. The second section provides a detailed analysis of the WWC's recent report on *Reading Mastery*, describing the extensive extant body of scientific research on the program and serious errors within the WWC report. The third section describes more general concerns with the work of the WWC and misrepresentations of research on one other curricular program, and the final section provides recommendations for changes in WWC procedures. Appendices provide supporting material.

The What Works Clearinghouse

The What Works Clearinghouse (WWC) was established in 2002 through a grant from the U.S. Department of Education's Institute of Education Sciences. Billing itself as "a central and trusted source of scientific evidence for what works in education," the Clearinghouse reviews studies of educational interventions, assessing the "rigor of research evidence" and "giving educators the tools to make informed decisions" (http://ies.ed.gov/ncee/wwc/aboutus/). The WWC has focused its work on analyzing individual curricula, giving its highest rankings to generally small, but tightly controlled, experimental studies on the impact of different curricula.[1]

After a very slow start, the Clearinghouse began issuing reviews in 2007, periodically posting summary judgments of individual curricular programs on its website. The assessments appear to be internally developed by WWC staff with no external oversight of the process or results. Unlike most of the scientific literature, the assessments are not exposed to peer review or other independent judgment by experts in the field. Not surprisingly, the work of the WWC has faced extensive criticism for its procedures and methodology including the ways in which studies are identified, the manner in which summary rankings are constructed, the accuracy of the reviews that have been produced, and the absence of any outside or independent oversight (e.g. McArthur 2008, Slavin, 2008).

The author of this document queried the WWC regarding its practices in e-mail communications from December, 2007 through April, 2008. Based on the responses received to those queries and further examination of WWC materials, a detailed letter was sent to Mathematica, the organization responsible for administering the Clearinginghouse on June 25, 2008. This letter summarized concerns with the WWC's procedures and policies. A reply was received from Mathematica on September 8, with a cover letter from the CEO of Mathematica and a more detailed report, apparently prepared by Mark Dynarski, the project director of the

[1] This focus is somewhat surprising given the emphasis within the educational community and research literature on the importance of comprehensive changes within schools to produce long-lasting changes in achievement. In addition, the WWC's "Statement of work" claims that the goal of the clearinghouse was to develop reviews of "educational interventions," parenthetically defined as "programs, products, practices, and policies" (WWC Statement of Work, Section I.B., paragraph 1).

WWC. All of these communications are included in Appendix A and are occasionally referenced in this document.

Unfortunately, none of the information that has been provided by the WWC or Mathematica has eased the concerns with the WWC's procedures that were outlined in the original queries. The material below summarizes some of these concerns and indicates why the conclusions of the WWC, both with respect to *Reading Mastery* and to other instructional programs, should be seriously questioned.

Direct Instruction, *Reading Mastery* and the WWC

Numerous studies have demonstrated the effectiveness of Direct Instruction with a wide variety of students in very different settings. Meta-analyses that examine large bodies of research literature also consistently point to the superiority of Direct Instruction. For instance, a recent meta-analysis of the achievement effects of different models (Borman et al 2003) found that Direct Instruction was the most researched of any of those studied and had the most research conducted by those not affiliated with the developers. As they put it, "the research base for Direct Instruction is very extensive and of very good quality" (p. 187). Among the many different models used in their analysis, Direct Instruction had the strongest evidence of effects on student achievement. Numerous other analyses have reached similar conclusions (e.g. Adams and Engelmann 1996, AFT 1998, Beck and McCaslin 1978, Herman, et al 1999, Kennedy 1978, Schieffer et al 2002). Additional studies expound upon the importance of the principles that underlie Direct Instruction, especially systematic and explicit instruction (see Adams 1990, Anderson et al 1985, Baker 1994, Bond and Dykstra 1967, Chall 1967, Foorman 1995, Fukkink and deGlopper 1998, Grossen 1997, ICHHD 2000, Juel and Minden-Cupp 2000, National Reading Panel 2000, NICHHD 1996, Pflaum, et al 1980, Smith et al 2001, Snider 1990, Snow et al 1998, Stanovich 1994). Finally, Direct Instruction has been lauded as not just the purveyor of effective curricular materials, but also as an important model of comprehensive school reform, guiding the transformation of schools into effective organizations in which all students can have high achievement (e.g. Borman 2002, Buechler 2002, CSRQ 2006).

Direct Instruction's *Reading Mastery* curriculum is a comprehensive core reading program with multiple levels and is the most widely used program within the DI curricular library. It has long been acclaimed as highly effective, with strong research accumulated over decades of work, supporting its development and documenting its success.

Despite the large body of research on Direct Instruction and the various associated curricular programs, *Reading Mastery* was not included in the initial WWC review process for Beginning Reading. Through mid-summer, 2008, the WWC website simply stated, without further explanation, that they chose not to include *Reading Mastery* in their review of Beginning Reading material.[2] In mid-August, 2008, however, the WWC issued a report on *Reading Mastery* with the following conclusion:

[2] In July, 2007, the WWC posted a review of *Corrective Reading* on its Beginning Reading page. *Corrective Reading* is a Direct Instruction program designed for children in the upper elementary grades who are reading below grade level. No explanation was given for why this program was included in the Beginning Reading category when its target audience did not match either the intended grade level or the general classroom focus of the review. While the review accepted one study of *Corrective Reading* as meeting its criteria and determined that it showed

> No studies of *Reading Mastery* that fell within the scope of the Beginning Reading review meet WWC evidence standards. The lack of studies meeting WWC evidence standards means that, at this time, the WWC is unable to draw any conclusions based on research about the effectiveness or ineffectiveness of *Reading Mastery*.[3]

To support their conclusion the WWC report lists 61 citations on *Reading Mastery* that were examined and a summary of the reasons they did not meet the WWC's standards of evidence.

Because the WWC conclusion regarding *Reading Mastery* contrasts so starkly with previous findings of the scientific community, the research staff of the National Institute for Direct Instruction (NIFDI) carefully examined the listed works and compared them to the extant literature and other reviews. As detailed below, our examination uncovered numerous examples of faulty judgments by the WWC. These involve three major points, all of which are elaborated below: 1) the WWC report ignored a large body of the extant research through seemingly arbitrary review decisions, 2) the WWC report did not find or review a significant proportion of the research on *Reading Mastery* that should have met their review criteria, and 3) the few reviews that were conducted often misconstrued and misinterpreted the research evidence.

Based on our review, others' critiques of the work of the WWC, and the very extensive research base that has documented the effectiveness of Direct Instruction and *Reading Mastery*, we conclude that the WWC statement regarding *Reading Mastery* is inaccurate and misleading. A large body of literature points to the very strong effect of Direct Instruction and *Reading Mastery* on children's reading achievement.

The WWC Ignored a Significant Body of Research on *Reading Mastery*

The WWC limited its review of studies for the Beginning Reading category to those that involve students in general education in grades K-3 and that were published between 1985 and 2007. When NIFDI's Director of Research questioned the decision to ignore studies published before 1985, the WWC responded that the "cut-off date of 20 years is a parameter set to ensure the research reviewed is most relevant to classrooms as they operate today" (e-mail correspondence, April 9, 2008, see Appendix A). The report included with the September 8 letter from Mathematica expands upon this reasoning by suggesting that numerous other changes over time such as increased levels of reading readiness (supposedly as a result of increased preschool enrollment), "stronger training" of teachers and the use of "newer curricula" "could have implications for the effectiveness of an intervention" (Dynarski, page 2). The exposition in the

---

"potentially positive effects," the issues raised in this document regarding WWC's review of *Reading Mastery* also apply to its review of *Corrective Reading*. One study of *Reading Mastery* was also included in the ELL portion of the WWC webpage, but again this placement does not reflect the target audience of *Reading Mastery*. It should be noted, however, that the high ratings these studies of both *CR* and *RM* received, even with these alternative audiences, should be seen as further evidence of the quality of the programs.

[3] The WWC Reading Mastery report, including the studies they reviewed, can be found at http://ies.ed.gov/ncee/wwc/pdf/WWC_ReadingMastery_081208.pdf. Interestingly, the September, 2008 letter from Mathematica (Dynarski report, page 4) notes the addition of the RM review to the website on August 12, 2008. It is unclear to what extent the review was prompted by our letter of June 25, 2008.

correspondence pointedly excludes any mention of empirical data to support the causal linkage that is suggested.[4]

We have not been able to find any studies cited on the WWC website, or elsewhere, that document that children's learning styles, classrooms, teachers' actions, or schools' organizational structures have drastically changed since 1985. As noted above, an extensive body of research cites the importance of systemic and explicit instruction in promoting achievement. This research spans the last 4 decades and has produced consistent results, no matter when the research was conducted (see Adams 1990, Anderson et al 1985, Baker 1994, Bond and Dykstra 1967, Chall 1967, Foorman 1995, Fukkink and deGlopper 1998, Grossen 1997, ICHHD 2000, Juel and Minden-Cupp 2000, National Reading Panel 2000, NICHHD 1996, Pflaum, et al 1980, Smith et al 2001, Snider 1990, Snow et al 1998, Stanovich 1994). In all the other areas of scientific study with which we are familiar, such as medicine, psychology, or the other social sciences, the full range of research articles are used to formulate policy, and there is no arbitrary date for inclusion or exclusion of results. If changes over time are hypothesized, the possibility of such changes is subjected to empirical test. Many analyses of the research literature systematically examine variations in results by the time when the study occurred.

This limitation of the time period for review had a disproportionately large impact on the number of studies of Direct Instruction included in the WWC's scope of study compared to other interventions precisely because DI has such a long and strong research base. A strong body of seminal research on DI was conducted before 1985, including Project Follow Through, the largest educational experiment in the history of the United States.[5] In general, research on interventions tends to be most active soon after the creation of the program. Because *Reading Mastery* was developed in the 1960s and high quality research conducted at that time demonstrated its effectiveness with a wide range of students, researchers have had little incentive to conduct additional studies. Given the consistently positive results from earlier decades, the current recent research focus has moved to populations of students with more extensive needs, such as older children who had not yet learned to read.[6]

A significant amount of research on *Reading Mastery* was ignored by the WWC because of the seemingly arbitrary choice of 1985 for a cut-off. Appendix B of this report includes a partial list of studies of *Reading Mastery* and its precursors that were completed before 1985.[7] The list was compiled by reviewing published literature reviews

[4] The citations provided are to macro-level data regarding increased preschool and kindergarten enrollment as well as to a study regarding the relationship of race and socio-economic status to achievement. None relate, either directly or indirectly, to variations over time in the impact of curricula or to changes over time in children's learning styles or achievement.

[5] Project Follow Through employed a strong experimental design and an extremely large sample to compare a wide variety of elementary curricula. The results provided very strong evidence of the superiority of Direct Instruction in raising academic achievement as well as students' self image.

[6] Studies of children in special education should have been within the purview of the review. In fact, the description of "populations to be included" in the WWC's Beginning Reading protocol states that "because students with learning disabilities…lag behind the population as a whole in reading achievement, studies involving these groups are of particular interest." No mention is made of excluding studies that involve students in special education.

[7] The cut-off date of 1985 used for the Reading Mastery report differs from the general WWC protocol for the Beginning Reading studies, which uses 1983 as the cut-off date. We have found no explanation for this discrepancy

and meta-analyses, all of which should have been available to the WWC staff. The list includes 38 items.[8] It should be stressed that this list is not exhaustive, but simply relied on easily available meta-analyses and reviews that should have been accessible to the WWC staff.

In short, unless the WWC can provide scientific evidence to the contrary, it would appear reasonable to suggest that high quality studies conducted before 1985 are still relevant today and should be included in any evaluation of *Reading Mastery*. Any conclusion regarding the effectiveness of *Reading Mastery* that does not consider this part of the literature should be challenged as incomplete and inaccurate.

The WWC Failed to Review Many Existing Studies of *Reading Mastery*

In addition to excluding studies published prior to 1985, the WWC review omitted a large number of studies published after 1985 that could have been included. Appendix C provides a partial list of these studies and includes 56 different citations. Like Appendix B, this list was largely generated by examining earlier reviews that should have been available to the WWC staff. Although the WWC lists a methodology for retrieving studies, the fact that they could miss so many citations that were in already published reviews suggests serious difficulties with their selection criteria.

As with the vast majority of work in this area, the studies listed in Appendix C generally found that *Reading Mastery* was superior to other approaches. For instance, an article by Kamps and associates (2003) uses multi-level analytic techniques to examine data from schools that implemented different reading curricula and found that students in *Reading Mastery* had significantly stronger achievement gains than in any of the other programs, both for those with and without social behavioral risks. Similarly, Gunn and associates (2000) report a study in which students were randomly assigned to interventions and found that both Hispanic and non-Hispanic students who received instruction in *Reading Mastery* had significantly more improvement over time than other students.

It should also be noted that the WWC list of writings on *Reading Mastery* includes material that would generally not be included in summaries of the scientific literature. Many items on the list are not research reports and were never intended as such. A large proportion is reports by school principals, teachers, and/or superintendents of the success their schools have experienced with Direct Instruction. These testimonials appeared in publications of the Association for Supervision and Curriculum Development and Council of Chief State School Officers (accounting for 8 of the items on the list of 61 studies) and on the website of SRA, the publisher of DI materials (26 of the items).[9] Two other studies on the list clearly targeted students outside the stated age range, with the titles indicating that they looked at "ninth and tenth graders" (Airhart 2005) or at "middle

[8] A large number of these studies examined DISTAR, the name by which the content now in Reading Mastery, was once known.

[9] Most of the testimonials include results from norm referenced achievement tests, often involving data from several years. These could have been used, if the WWC desired, to compare the achievement in these schools with national norms.

school students" (Shippen et al 2006). An additional study (Flores and Ganz 2007) did not look at *Reading Mastery*, but at *Corrective Reading*, a different Direct Instruction curricular program. Three studies in the list did not have reading achievement as a dependent measure, but instead were content analyses of textbook stories (Jordan 2005), a qualitative study of three students' understandings of what they read (Wilson 2005), and an examination of students' social behaviors (Smolkowski et al 2005). Finally, a few others appeared to have never been subjected to the usual scientific standards of peer review, such as articles posted on websites, papers given at professional meetings, and unpublished dissertations (6 studies in all).[10, 11]

When these 46 studies are omitted from the list, the WWC actually examined only 15 studies of *Reading Mastery* that would generally be considered part of the scientific research base. This number is less than one-fourth of the number given in the report. The listing of materials in the WWC's report on *Reading Mastery* implies that the research base that was reviewed was much larger than it, in fact, was.

In general, it appears that the WWC list of materials that were reviewed suffers from errors of both exclusion and inclusion. A large number of studies appear to have been ignored. Our review of already existing reviews of the literature revealed almost one hundred articles that could have been consulted if a full analysis were to occur. These citations were easily found through simply searching the reference list of already published literature reviews and thus are no doubt an undercount of the available resources. The WWC review missed a large proportion of the research base, not just through the time parameters that were used but also, apparently, through a search of the literature that was far from thorough.[12] At the same time, much of the material listed in their report should have fallen outside the domain of a comprehensive review of the

[10]Meta-analyses often include such unpublished works in their reviews, and that has been the case with several meta-analyses of beginning reading. Given that the WWC review is not subject to peer review, omitting these studies from their work could be one way to add additional controls. If the WWC's work were subject to peer review, including unpublished work would be less problematic. Dynarski's report included with the September, 2008, letter addresses the issue of peer review, citing a study of publication bias in medical research. Whether or not the result with medical research applies to educational research is an empirical question and a citation to the education or more general social science literature would be more appropriate. We know of no study that has found that high quality social science work remains unpublished. In any case, the standard practice in meta-analyses and other reviews is to treat the issue of the relationship of publication status to results as an empirical question and compare the unpublished and published articles.

[11] The WWC categorized one of the studies that had not been subjected to peer review (CSRQ 2006) as "ineligible…because it does not examine the effectiveness of an intervention." This study, prepared by the highly regarded Comprehensive School Reform Quality Center, reviewed and summarized the evidence on the effectiveness of 22 different school reform models, including Direct Instruction. As with other such reviews, there were more studies available regarding the effectiveness of Direct Instruction than for any other model and the results indicated that Direct Instruction was consistently one of the most effective, if not the most effective, in each of the categories examined. Thus, while the CSRQ document did not address the effectiveness of a single intervention, to dismiss it because it examined numerous models could be seen as misleading.

[12] Ironically, one of the literature reviews used to generate the listings in Appendices B and C was the 2006 report of the Comprehensive School Reform Quality Center (source e in the listings) and discussed in footnote 10 above. The WWC lists this report as one that they consulted and found "ineligible for review," but apparently did not use its references for further examination.

scientific literature, providing a misleading representation of the amount of literature that was examined.

Inaccuracies in the Reviews of *Reading Mastery* Studies

The WWC report provides one sentence summaries of the reason that a study was judged not to meet their established "evidence standards." These reasons involved either questions about the comparability of the control and experimental groups or the existence of a possible "confound" to the treatment. Examples are given below of the studies rejected for these reasons and explanations of the faulty logic involved in the conclusions. The errors in these examples are so serious that they should raise serious concerns regarding the WWC's evaluation procedures.

Equivalence of Experimental and Control Groups – Three of the published studies examined by the WWC that addressed student achievement were rejected because, 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."[13] The NIFDI research staff carefully examined these three articles and found that, in each case, the conclusion of the WWC could be questioned. In one of the studies (Brent et al, 1986) the presentation of the results clearly noted that adjustments had been made, using standard multivariate procedures, for any possible pretest differences between the treatment and control conditions. In the two other studies (O'Brien and Ware 2002, Thomson 1991), the pretest scores of the treatment group (those receiving *Reading Mastery*) were lower than the scores of students in the control group. Such a situation is routinely considered as an indication of a conservative test, for it biases results against the treatment, and is not used as a reason to automatically discount the results. Again, supporting the large body of extant evidence, these three studies all showed positive effects of *Reading Mastery* on students' achievement.

Confounding Factors – Nine studies were rejected from the WWC's review because, although they used a quasi-experimental design, the reviewers determined that there was "a confounding factor, such as combining with other interventions, which makes it impossible to attribute the observed effect solely to *Reading Mastery*." Unfortunately, the WWC provided no specific details in the report published on the web regarding how they reached this conclusion for each study. Again, however, the review by the NIFDI Research staff suggests that this judgment may not be appropriate for at least some of the works.

For instance, one of the studies rejected in the *Reading Mastery* review for having a "confounding factor" was Carlson and Francis's (2002) evaluation of the RITE program. For some reason this study had mistakenly been included in WWC's earlier review of the Direct Instruction *Corrective Reading* program and had also been rejected in that review for a supposed confound. This rejection was questioned in the letter written to Mathematica on June 25, 2008, and the response to this query in the September 8 correspondence sheds greater light on the WWC's definition of a "confound." This explanation also, however, raises serious questions about the validity of the WWC review and its implications for educators.

[13] The WWC report lists 6 studies that were rejected for this reason, but three of these were unpublished works such as those described above.

The Carlson and Francis study involved thousands of students, hundreds of teachers, and dozens of schools and used sophisticated multivariate techniques to analyze the data. The only intervention in the experimental schools was Direct Instruction's *Reading Mastery*, and as with the other studies, the results strongly favored this curriculum. After scouring the article, the NIFDI research staff found one possible explanation for the WWC's determination. On page 143, the authors described the intervention: "In addition to the teaching of skills directly related to the RM curricula, the … program also strives to provide teachers with strong classroom management techniques" (p. 143).[14] The September 8, 2008, communication from Mathematica confirms this interpretation, with the statement that, "A careful reading of Carlson and Francis indicates that findings cannot be separated into effects of *Reading Mastery* alone and effects of *Reading Mastery* supplemented by the support provided to teachers through the RITE program" (Dynarski, p. 5).

In reality, strong classroom management is part and parcel of the Direct Instruction approach. Training teachers in such management is part of the in-service training that teachers receive in learning how to implement the curriculum as well as prominently included within the teacher's guide to the program. Part of the reason that DI is so successful is that it provides not just well designed curricular materials but well developed, research-based guidance on how the curriculum should be administered. Because these guidelines are well documented in the various guides to the programs, they should have been part of the knowledge base of a competent reviewer. The classroom management was not a confounding element of the intervention, but was an integral part of the curriculum and its appropriate delivery.

More importantly, virtually all curricular programs include elements of teacher training and discussions of classroom management. It is reasonable to argue that if studies of *Reading Mastery* that include training for teachers are to be excluded, all other studies that include training for teachers should also be excluded. We know of virtually no legitimate curriculum that does not include some type of instructional overview for teachers.

In addition, complete training in a program is vital to ensuring that it is implemented with fidelity. Disallowing such elements from a design could produce a very serious threat to internal validity of a study. Surprisingly, the WWC's approach seems to discount the importance of the fidelity of treatment implementation, suggesting that "there is no standard metric" with which to rate and assess fidelity" and that a better approach is their reliance on "replicated findings, which ensures that any one study in which fidelity issues may have arisen are averaged with findings from other studies" (Dynarski 2008, p. 3).

At least two very serious problems are immediately apparent with these statements. First, those familiar with any number of structured learning and behavioral approaches know that there is in fact a well developed literature of fidelity measures and that such measurements are very important both in helping to train practitioners and to assess the extent to which programs are being adequately implemented. To suggest that such measures do not exist is to demonstrate an unfortunate unfamiliarity with the literature. Second, to rely on replications as a substitute for

[14] The WWC Reading Mastery bibliography lists the Carlson and Francis article as in the *Journal of Direct Instruction*. The article also appeared in the *Journal of Education for Students Placed at Risk* in 2002, volume 2, issue 2, pp. 141-166. The page number in the text refers to this citation.

measures of fidelity could well produce very misleading results. Numerous replications of a poorly implemented program can, in no stretch of the imagination, provide an adequate test of a program. To assume that poor implementations are so rare that their impact would "average" out is, at best, a contention that should be subject to empirical test. More realistically, relying on such an assumption would likely produce very inaccurate and misleading results, for the errors that result from ignoring fidelity are systematically biased. As detailed in Appendix D, including studies with poor fidelity within summary analyses produces higher ratings of poor programs and lower ratings of good programs, thus providing very misleading conclusions for consumers.

Finally, two other studies that were rejected for an alleged confound were those by Gunn and associates (2002, 2005). These works report the results of a study that included both a behavioral intervention and *Reading Mastery*. While it is understandable that a naïve reviewer might quickly reject these articles, a careful reading shows that the authors directly address the issue of any possible confound and, based on both their results and other works in the field, note that this was unlikely. Again, these articles, plus a third by the authors that was not included in the WWC listing, find results strongly in favor of *Reading Mastery*.[15]

Summary

To summarize, examination of the What Works Clearinghouse's review of *Reading Mastery* indicates a number of aspects that should lead consumers to question the accuracy of the report. Part of the problem involves the ways in which studies were selected for inclusion and exclusion from review. An arbitrary time frame resulted in a large number of studies being excluded. In addition, however, the Clearinghouse ignored an even larger number of studies published within the chosen time frame. It appears that the WWC made no attempt to examine any of the large and easily available meta-analyses and literature reviews and use the studies listed in those documents. In addition, the accuracy of the WWC report should be questioned because of the quality of the reviews. The reviews appear to be both inaccurate and misleading.

Inaccuracies in Other WWC Reports

As noted above, the letter of June 25, 2008 to Mathematica was written before the *Reading Mastery* review was posted. One issue discussed in the letter is of such concern that it deserves additional discussion in this comment: the misrepresentation of results for at least one program that was given high ratings.

The WWC has given high ratings to the Reading Recovery program, a short-term tutoring intervention. The WWC website concludes that "*Reading Recovery*® was found to have positive

[15] Of the remaining three peer reviewed studies examined by WWC, two were excluded from review because they did not include a comparison group, but relied on comparison of pretest and posttest scores for a single group of subjects (Humphries et al 2005 and Marchand-Martella et al 2006). A third (Kamps and Greenwood 2005) was excluded because it "does not examine the effectiveness of an intervention." Examination of this study indicates that, in fact, just the opposite is true – the study reports the results of the first few months of a randomized trial of interventions comparing several more highly structured curricula, including Reading Mastery, with less structured approaches. Although descriptive statistics indicate that the students exposed to Reading Mastery have had substantially higher rates of progress than students in the traditional settings, inferential statistics that examine differences between the settings were not included. Thus, while definitive judgments could not be made regarding the statistical significance of the results, to indicate that this study does not involve an intervention is, at best, misleading.

effects on students' alphabetics skills and general reading achievement outcomes and potentially positive effects on comprehension and fluency" (http://ies.ed.gov/ncee/wwc/reports/beginning_reading/reading_recovery/). A careful reading of the articles cited by the website, however, indicated that, in at least two cases, the WWC review did not accurately reflect the content of the research studies.

For instance, one article (Baenen, et al 1997) was cited as showing positive impacts of *RR*, yet even a cursory reading of the article indicates that the authors found remarkably little success, especially in the long-term, for students in the program. The article concludes (p. 176) that although about one-half of the students in the tutoring program had successfully reached first grade reading levels at the end of the year, success rates declined in subsequent years of implementation. In addition, by third grade, there was no difference in achievement scores, needs for retention, special education, or Chapter 1 assignments of students who had participated in *Reading Recovery* and other students. The authors concluded that *Reading Recovery* was very expensive to implement in relation to the benefits that it provided.

The seemingly erroneous conclusion of the WWC was brought to the attention of Mathematica in the June 25 letter. The response (Dynarski 2008, p. 4) acknowledged the results outlined above, but defended the WWC's conclusions 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. Yet, as teachers, parents and students would attest, how well one reads at the end of third grade is much more crucial in determining eventual success than how well one reads at the end of first grade. Learning to read is not a one-year process, but is a multi-year endeavor. If gains in first grade do not persist through third grade, the first grade gains have very little worth. Ignoring the very important conclusions of Baenen and associates would seem to do a disservice to parents and teachers.

An even more disturbing example involves an article by Iversen and Tunmer (1993), which was also cited by the WWC as supporting the conclusion that *RR* is effective. The major purpose of this study was to compare the standard *RR* program to a "modified" program that included explicit instruction in phonological skills. 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 was found to be 37 percent less efficient than the modified program. In addition students in the modified program continued to have higher levels of achievement and higher rates of learning at the end of the school year. The authors provide an extensive discussion and additional analyses that demonstrate the fallacy involved in *Reading Recovery* about the ways in which word recognition skills develop. They clearly conclude that *Reading Recovery* is not an efficient method for teaching children to read and that phonological training is superior.

The WWC chose to ignore any results regarding this comparison group, which received phonological training and had superior achievement, "because it was a modified version of the

standard program." The September response to our query regarding their judgment stated, "the WWC examined the results most relevant to the question of whether Reading Recovery improves reading proficiency compared to a reasonable counterfactual." That counterfactual was, apparently, having no tutoring at all, not tutoring with a more effective program. The statement goes on to note that the results with the other comparison groups were mentioned in an appendix, implying that such information could be available for those who were interested.

Even though the WWC did acknowledge the negative results regarding *Reading Recovery* in a technical Appendix, the overall conclusion given on its website is certainly misleading and does not reflect the conclusions of the articles. The chance of a parent or school official accessing a technical appendix to find information that contradicts that given in the major pages of the web site is extremely remote. The actions of the WWC regarding the material on *Reading Recovery* not only resulted in misrepresenting the effectiveness of *Reading Recovery* but also that of phonologically based programs and provides further evidence that its conclusions should be seriously questioned.

Recommendations

Based on a careful review of the evidence, we suggest that consumers approach the ratings of the WWC with extreme skepticism. With respect to *Reading Mastery* the list of sources that were reviewed was extremely selective and tapped only a very small proportion of the extant literature. Perhaps even more disturbing, the reviews of the works that were selected were often inaccurate and misleading. We see no reason to accept their conclusion of "no evidence" as valid.

From our correspondence with the WWC there also appear to be a number of very serious policy and procedural issues that raise questions regarding the accuracy of any ratings that they have developed. As detailed above, their policies in a number of areas appear to differ substantially from traditional scientific practices. These include such areas as decisions about how to select studies to examine, classifying teacher training procedures as a confounding element, relying on replications to "average out" problems with fidelity of implementations, and focusing on one-year studies rather than multi-year results, even when such multi-year results are available and more indicative of student success.

Part of the reason that the WWC has reached erroneous conclusions may reflect the way in which its review process departs from well established traditions of scientific research. In our June, 2008, letter to *Mathematica*, we made several suggestions for changes to their procedures. These suggestions in many cases overlapped with those made by other critics of WWC (see Slavin 2008 and McArthur 2008). For instance, we suggested that using standard quality control mechanisms, such as two reviewers for every article with a third if results differ and a peer review process before publication of ratings, could go far to help ensure more accurate results. We suggested that the cumulative nature of the scientific enterprise should be acknowledged, and the WWC's results should be compared with the already well developed body of meta-analyses and literature reviews. When WWC's results differ from these reviews, they should investigate why and adjust the results as needed. In addition, metrics that are commonly used within the social sciences, such as measures of effect size would greatly enhance faith in the WWC's conclusions. Issues of both internal and external validity in decisions regarding acceptable

designs should be weighed. Much more accurate information could be provided if the WWC understood the wide variety of research approaches that can and must be used within real-life educational settings. 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. Finally, it is crucial that the WWC ensure that reviewers are knowledgeable in the substantive areas that they are reviewing as well as in the methodological details.

Unfortunately, the concerns expressed in the correspondence with Mathematica and the WWC, as well as the concerns of earlier critics Slavin (2008) and McArthur (2008) appear to have been unheeded and, in fact, rejected. As stated in that letter, the recommendations were made "in the spirit of the intent behind the What Works Clearinghouse – a desire to provide schools and families with the most accurate information." The letter also noted that "the problems with the material posted on the website appear to be much more widespread than can be handled in a piece-meal fashion. To allow inaccurate material to remain does a disservice to schools, families and students."

The communications with Mathematica appeared before the review of *Reading Mastery* was posted. Unfortunately, the quality of that review has made the earlier judgments of problems with the WWC even more apparent and worthy of grave concern. If Mathematica refuses to remove the faulty information from the WWC website, consumers should be extremely wary of the ratings posted there and consult the standard social science literature for more accurate information.

References

Adams, Gary L. and Siegfried Engelmann. 1996. *Research on Direct Instruction: 25 years beyond DISTAR*. Seattle: Educational Achievement Systems.

Adams, Marilyn J. 1990. *Beginning to read: Thinking and learning about print*. Cambridge, Mass.: The MIT Press.

American Federation of Teachers. 1998. *Building on the best, learning from what works: Seven promising reading and language arts programs*. Washington, D.C.: AFT.

Anderson, R. C., E. H. Hiebert, J. A. Scott, and I. A. G. Wilkinson. 1985. *Becoming a nation of readers: The report of the Commission on Reading*. Washington, D.C.: National Institute of Education.

Baenen, Nancy, Alissa Bernholc, Chuck Dulaney, and Karen Banks. 1997. Reading Recovery: Long-Term Progress After Three Cohorts. *Journal of Education for Students Placed at Risk* 2: 161-181.

Baker, Scott K. Edward J. Kameenui, Deborah C. Simmons, and S. A. Stahl. 1994. Beginning reading: Educational tools for diverse learners. *School Psychology Review* 23: 372-391.

Beck, I.L. and E. S. McCaslin. 1978. *An analysis of dimensions that affect the development of code-breaking ability in eight beginning reading programs*. LRDC Report No. 1978/6 Pittsburgh.

Bond, G. and R. Dykstra. 1967. The cooperative research program in first-grade reading instruction. *Reading Research Quarterly* 2: 5-142.

Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2003. Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research* 73: 125-230.

Buechler, Mark. 2002. *Catalog of School Reform Models: Program Report*. Portland, Oregon: Northwest Educational Research Lab.

Chall, J.S. 1967. *Learning to read: The great debate*. New York: McGraw Hill.

Comprehensive School Reform Quality Center. 2006. *CSRQ Center Report on Elementary School Comprehensive School Reform Models*. Washington, D.C. American Institutes for Research.

Dynarski, Mark. 2008. "Responses to Concerns in 6/26/08 Letter from Jean Stockard," sent on September 8, 2008, from Mathematica, included in Appendix A of this report.

Foorman, B. R. 1995. Research on "the great debate": Code-oriented versus whole language approaches to reading instruction. *School Psychology Review* 24: 376-392.

Fukkink, R.G. and K. deGlopper. 1998. Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research* 68: 450-469.

Grossen, B. 1997. *A synthesis of research on reading from the National Institute of Child Health and Human Development*. Eugene, Oregon: University of Oregon.

Herman, R., D. Aladjam, P. McMahon, E. Masem, I. Mulligan, O. Smith, A. O'Malley, S. Quinones, A. Reeve, and D. Woodruff. 1999. *An educator's guide to schoolwide reform*. Washington, D.C.: American Institutes for Research.

Institute of Child Health and Human Development (ICHHD). 2000. *Report of the National Reading Panel. Teaching Children to Read: An Evidence-Based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, D. C.: U. S. Government Printing Office.

Iversen, Sandra and William E. Tunmer. 1993. Phonological processing skills and the reading recovery program. *Journal of Educational Psychology* 85: 112-126.

Juel, C. and C. Minden-Cupp. 2000. Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly* 35: 458-492.

Kennedy, M.M. 1978. *Findings from the Follow Through Planned Variation Study*. Washington, D.C.: U.S. Office of Education.

McArthur, Genevieve. 2008. Does what works clearinghouse work? A brief review of FastForWord®. *Australasian Journal of Special Education* 32: 101-107.

National Institute of Child Health and Human Development (NICHHD). 1996. Thirty years of NICHD research: What we now know about how children learn to read. *Effective School Practices* 15: 33-46.

National Reading Panel. 2000. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, D.C.: National Institute of Child Health and Human Development.

Pflaum, S., H. J. Walberg, M. L. Karigianes, and S.P. Rasher. 1980. Reading instruction: A quantitative analysis. *Educational Researcher* 9: 12-18.

Schieffer, Cheryl, Nancy E. Marchand-Martella, Ronald C. Martella, Flint L. Simonsen, and Kathleen M. Waldron-Soler. 2002. An analysis of the *reading mastery* program: Effective components and research review. *Journal of Direct Instruction* 2: 87-199.

Slavin, Robert E. 2008. Perspectives on evidence-based research in education – What works? Issues in synthesizing educational program evaluations. *Educational Researcher* 37: 5-14.

Smith, S., D. Simmons, M. Gleason, E. Kameenui, S. Baker, M. Sprick, B. Gunn, and C. Thomas. 2001. An analysis of phonological awareness instruction in four kindergarten basal reading programs. *Reading and Writing Quarterly* 17: 25-50.

Snider, V.E. 1990. Direct instruction reading with average first graders. *Reading Improvement* 27: 143-148.

Snow, C. E., M. S. Burns, and P. Griffin (eds.). 1998. *Preventing reading difficulties in young children*. Washington, D. C.: National Academy Press.

Stanovich, K.E. 1994. Romance and reality. *The Reading Teacher* 47: 280-291.

Appendix A
Communications with the What Works Clearinghouse and Mathematica

1) e-mail –December 2, 2007, from J. S. to WWC
2) Reply from WWC, February 8, 2008
3) JS to WWC, March 16, 2008
4) WWC to JS, April 9, 2008
5) JS to Mathematica, June 25, 2008 (letter sent by U.S. post)
6) Follow-up letter JS to Mathematica, July 31, 2008 (sent fed-ex with receipt notice)
7) Reply from PD to JS, August 6, 2008
8) JS reply to PD, August 7, 2008
9) A. S. to JS, September 6, 2008
10) Reply from Mathematica, September 8, 2008

1) e-mail –December 2, 2007, from J. S. to WWC

Sent: Sunday, December 02, 2007 12:30 PM

To: Mark Dynarski

Subject: Questions on What Works Clearing House Procedures

Dear Dr. Dynarski:

I am a sociologist with many years of experience in quantitative research and recently began working with the National Institute for Direct Instruction (NIFDI). One of my first tasks was to read as much of the background research in the area as I could. Along the way I ran into the "What Works Clearinghouse." In my review I developed several questions that I am hoping you can answer.

First, I noticed that the WWC has listed relatively few of the studies that deal with Direct Instruction. While a few were rejected because they didn't meet some specific standards of design, it appears that many others were rejected because they were published before 1985. I have not been able to find a scientific justification for selecting this date. No studies were cited that suggested children's learning styles had altered after that time or that schools' organizational structures or teachers' actions had changed. In fact, the only research article that I found to justify the methods was actually published in 1974 – 10 years prior to the beginning of your targeted dates. This article dealt with phonological awareness, and had nothing to do with systematic literature reviews and, of course, included nothing to suggest that processes related to phonological awareness have changed over time.

In all the other areas with which I am familiar we use the full range of research articles and, unless there are clear reasons established through the research literature, do not set an arbitrary date for inclusion or exclusion of results. The choice of this date seems to have particularly affected the DI literature because a great deal of the pioneering work was published before the cut-off point. Because I'm new to this particular work, I wondered if you could provide me with any explanation based in the research literature for your choice of this cut-off year. I did find a suggestion that, "if sufficient time and resources remain," studies published before 1985 might be reviewed. Do you know if this step will be taken? From my reading of the literature I believe that your limit on dates of studies may have seriously limited the selection of effective resources that educators will find on your website; and I know that this is a primary aim of your project.

Second, I noticed that a number of studies were rejected because they included children outside the K-3 grade range. I found this somewhat unusual, for the children that are most in need of effective instruction are those who have not learned to read by older ages. Because the effectiveness of DI has been so well demonstrated with regular students at the K-3 grade range, the more recent focus on other ages seems only greater evidence of its worth. This is especially

so in the case of special education students and those with behavioral and other issues. Thus, I couldn't understand the rationale of omitting all the studies that targeted these most needy students. Any explanation you might have of this decision would also be greatly appreciated.

Third, and perhaps most important, as I read studies more closely I found that the Clearinghouse's interpretation of some of the studies didn't seem to accurately reflect the actual content of the reports. I know that these reviews are often done by staff members or graduate students who might not be familiar with all of the intricacies of social research. Thus, I assume that we can send corrections and they will be added.

I look forward to hearing from you, either through e-mail (jeans@uoregon.edu) or by phone (541-346-5005). I'm sure that we share the ultimate goal of helping all students – no matter what their background – develop academic skills and appreciate your appreciate your help.

Sincerely,
Jean Stockard, Professor Emerita

2) Reply from WWC February 8, 2008

Dear Dr. Stockard,

Thank you for contacting the What Works Clearinghouse (WWC). Please see the response to the questions from your email dated December 2, 2007 below.

The first question you raised concerned the cut-off year for studies that are reviewed by the WWC. To maintain a focus on current research, our reviews examine research conducted in the past 20 years. This cut-off date of 20 years is the standard for every topic reviewed by the WWC. The review for Early Childhood Education (ECE), under which Direct Instruction falls, marked its cut-off date at 1985 (the review began in 2005).

The ECE protocol states that "if sufficient time and resources remain after we have completed our review of research on interventions implemented post 1985, the ECE team will consider reviewing older research on curricula that are still in widespread use." This determination has not been made to date.

In terms of the age/grade range for the review of Direct Instruction research, the ECE team determined the parameters for the review. For further details on inclusion criteria please see the ECE protocol at http://ies.ed.gov/ncee/wwc/reports/early_ed/index.asp

In response to your third question about the qualifications of WWC topic review team members, each team is comprised of several trained professionals, each playing an integral part in the review process. The Principal Investigator (PI) for each topic is a well-known expert in his/her field and is responsible for leadership in conceptualizing the specific topic area, identifying and addressing issues during the review, and developing and reviewing topic and intervention reports developed for the topic. Leadership includes overseeing the quality in the production of the reports and making decisions, based on methodological and substantive expertise, that are not otherwise covered in the WWC protocols and procedures for report production.

The Project Coordinator (PC) is an established education researcher with relevant methodological and substantive expertise. The coordinator oversees the work of the WWC Review Team; manages that specific review; reviews research ratings; and writes and revises the work plan, protocol, and draft and final reports in collaboration with the PI.

Individual reviewers, who prepare initial summaries of studies for the WWC, are professional researchers with experience in research design and methodology. These reviewers undergo a rigorous training and certification process before conducting WWC reviews. .For more information about the staff that make up the WWC team, visit http://ies.ed.gov/ncee/wwc/overview/index.asp.

The WWC recently created a Quality Review Team to respond to concerns raised by study authors, curriculum developers or other relevant parties about WWC reviews published on our

website. These quality reviews are undertaken when concerned parties present evidence that a WWC review may be inaccurate. When a quality review is conducted, a researcher who was not involved in the initial review undertakes an independent assessment of the study in question. The researcher also investigates the procedures used and decisions made during the original review of the study. If a quality review concludes that the original review was flawed, a revision will be published. These quality reviews are one of tools used to ensure that the standards established by the Institute of Educational Sciences (IES) are upheld on every review conducted by the What Works Clearinghouse.

If you have concerns about a published WWC review that you think warrant a quality review, please send those concerns to our Help Desk at info@whatworks.ed.gov. Please identify the study in question, the specific issue(s) that you think were handled incorrectly, and where relevant, explain what you think is the correct interpretation of the study.

I hope you found this information helpful. If you have additional questions, please feel free to contact us again.

What Works Clearinghouse

The What Works Clearinghouse was established by the U.S. Department of Education's Institute of Education Sciences to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education. For more information, please visit http://ies.ed.gov/ncee/wwc/.

3) March 16, 2008 – JS to WWC

From: Jean Stockard [mailto:jeans@uoregon.edu]
Sent: Sunday, March 16, 2008 11:59 AM
To: WhatWorks
Subject: re: Questions on What Works Clearing House Procedures

Thank you for your reply of February 8 to my query of December 2. I am pleased to hear that you have established a Quality Review Team. Since my initial query I have continued to examine the reviews and have found quite a few that appear very problematic. I, and my colleagues, will be submitting our concerns to you within the coming weeks.

Unfortunately, I'm afraid that your reply failed to address two of my concerns. Your responses simply restated WWC's policies regarding the cut-off year for studies and the age-grade parameters without providing the requested justification. Again, I would be very interested in knowing the scientific and pedagogical justifications for not incorporating the entire corpus of work in your reviews and for limiting your reviews of studies to only those that included children in grades K-3.

My earlier, rather lengthy, explanation of my concerns is included in this e-mail. If you wish me to clarify any part of my query, please let me know. Thank you for your attention to this matter.

Sincerely,

Jean Stockard, Ph.D.
Emerita Professor

4) WWC to JS, April 9, 2008

Dear Dr. Stockard,

Thank you for contacting the What Works Clearinghouse (WWC). Please see the response to the questions from your email dated March 16, 2008 below.

In response to your first concern, the cut-off date of 20 years is a parameter set to ensure the research reviewed is most-relevant to classrooms as they operate today, and to ensure that the scope of studies to be reviewed is manageable. That said, for every topic reviewed by the WWC, the Principal Investigator (PI) is given the flexibility to include studies published earlier, if they think the expansion is important for the review. In this case, the decision was made to keep the 20 year parameter in place.

In terms of the age-grade parameters established, this is determined by the PI and depends on the topic area and the studies under review. In this case, because the focus of the topic is on early childhood education, the age-grade parameter was set to K-3. Other topic areas have also focused on a subset of age-grade ranges. As the WWC expands, we anticipate expanding these topics to examine outcomes for other age-grade ranges.

The WWC solicited nominations from many sources for the topic areas and prioritized the topic areas based on the following criteria:

- potential to improve important student outcomes;
- applicability to a broad range of students or to particularly important subpopulations;
- policy relevance and perceived demand within the education community; and
- likely availability of scientific studies.

More information about this process and the specific topic areas can be found at http://ies.ed.gov/ncee/wwc/reports/.

Three Direct Instruction interventions were reviewed in two topic areas; the spread across topics was due to the samples included in the studies. Direct Instruction falls under the Early Childhood Education (ECE) area since it includes studies with preschool and kindergarten children where the majority (60% or more) of children in the sample are in preschool. This criterion for inclusion can be found at http://ies.ed.gov/ncee/wwc/pdf/ECE_protocol.pdf. One Direct Instruction intervention is also included in the English Language Learners topic area since it was used to supplement reading instruction for Spanish speaking students in grades K-3.

We hope you found this information helpful. If you have additional questions, please feel free to contact us again.

What Works Clearinghouse

The What Works Clearinghouse was established by the U.S. Department of Education's Institute of Education Sciences to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education. For more information, please visit http://ies.ed.gov/ncee/wwc/.

5) JS to Mathematica, June 25, 2008 (letter sent by U.S. Post)

June 25, 2008

Anita A. Summers, Chairperson and
Paul T. Decker, President and CEO
Mathematica, Policy Research Inc.
P.O. Box 2393
Princeton, NJ 08543-2393

Dear Drs. Summers and Decker:

I write to express my deep concerns with the material posted on the What Works Clearinghouse, which was established "to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education." I understand that Mathematica has now assumed responsibility for the Clearinghouse. I also know that Mathematica has a well established reputation for high quality research. I write to you in the hope that this correspondence can help you correct the errors in the current reports and establish future review procedures that are in accordance with standard scientific practices. The goals of the What Works Clearinghouse are too important to allow a flawed review process to continue or faulty reviews to be posted. None of us wants parents or schools to be given faulty information about the most effective curriculum for their students.

I am a sociologist with many years of experience in quantitative research and recently began working with the National Institute for Direct Instruction. In beginning to familiarize myself with the area I, of course, reviewed major analyses of Direct Instruction. As you no doubt know, in recent years, meta-analyses have become the most commonly accepted method for examining large bodies of research literature. The most recent meta-analysis of the achievement effects of comprehensive school reform models was conducted by Geoffrey Borman and associates (2002, *Review of Educational Research*). Their examination found that Direct Instruction was the most researched of any of the models and had the most studies conducted by third parties (those not affiliated with the developers). As they put it, "the research base for Direct Instruction is very extensive and of very good quality" (p. 187). Among the many different models used in their analysis, Direct Instruction had the strongest evidence of effects on student achievement.

In my reviews I found that this conclusion simply echoes that found in earlier meta-analyses and literature reviews. The overwhelming conclusion of the education community for many years, based on solid experimental evidence, is that Direct Instruction is one of, if not the, most effective curriculum currently available for teaching reading and mathematics.

In my examination of the literature I also reviewed the reports on Direct Instruction that had been prepared by the What Works Clearinghouse and was shocked and dismayed to find that the WWC conclusions were in marked contrast to those of the extant literature. My concerns with what I found prompted me to write to the Clearinghouse on two previous occasions (December 2, 2007 and March 16, 2008). The replies I received did not clearly address my

concerns. Even worse, my continuing examination of the work of the Clearinghouse has raised even more serious concerns about the quality of the reviews.

In recent months others have expressed similar concerns regarding the Clearinghouse's conclusions and procedures. Two examples of these critiques are Robert Slavin's article in the January/February, 2008 issue of *Educational Researcher* and Genevieve McArthur's article in the April, 2008 issue of the *Australasian Journal of Special Education*. While Slavin and McArthur focused their concerns on reviews related to Success for All (Slavin) and *FastForWord*® (McArthur), many of the concerns they raise parallel the issues I discovered in my review of the judgments of articles regarding Direct Instruction.

Below I outline in greater detail the extent of my concerns. I first discuss what I see as severe limitations of the process used to decide what studies to include in the review process and then give examples of very severe errors in the reviews themselves. I end with a few suggestions regarding what needs to be done to provide the most accurate information to the education community and to parents.

Problems with Exclusion/Inclusion Decisions

Some of the most disturbing elements of the WWC process involve the ways in which studies are selected for reviews. Some of these decisions arbitrarily limit the range of studies included and thus the information available to schools and parents. For instance, as I noted in my earlier communications, the decision to reject all studies for inclusion that were published before 1985 appears to have no scientific justification. The reply to my April 9 query claimed that the "cut-off date of 20 years is a parameter set to ensure the research reviewed is most relevant to classrooms as they operate today." I have not been able to find any studies cited on the website, or elsewhere, that document that children's learning styles, classrooms, teachers' actions, or schools' organizational structures have drastically changed over that time period. Most high quality studies conducted before 1985 are still relevant today. As I stated in earlier correspondence, in all the other areas with which I am familiar we use the full range of research articles and do not set an arbitrary date for inclusion or exclusion of results.

Similarly, some areas of the review have a very restrictive grade range for studies that will be included. For instance, for reviews of work on beginning reading, a number of studies, of both Direct Instruction and other approaches as well, were rejected because they included children outside the K-3 range, such as grades 3-5 or grades K-4. It is hard to understand why such studies would be summarily rejected.

The combination of the decision to limit reviews to work published after 1985 and to a narrow band of grades directly affected the numbers of studies of Direct Instruction that were reviewed. Even worse, this decision ignored the cumulative nature of science and research inquiry within education. Many studies of Direct Instruction with only K-3 students were completed before 1985. Given the consistently positive results from that work, the research focus then moved to populations of students with more extensive needs, such as older children who had not yet learned to read. Yet these studies were omitted because they were not focused on the general population. The real losers from these decisions are, of course, children and their

families who have been denied accurate and complete information, collected over a long span of time and from many different populations, about the most effective programs.

Some of the decisions regarding exclusion or inclusion of curricula and studies seem to have occurred with no stated justification. For instance, Direct Instruction's *Reading Mastery* program has long been acclaimed as a highly effective reading program, with strong research supporting its development and documenting its success. Yet, the WWC website states that they chose not to include *Reading Mastery* in their review of Beginning Reading studies (apparently not realizing that the RITE study noted below involved *Reading Mastery*). No explanation is given as to why this curriculum, which directly addresses beginning reading and which has a very large research base, was ignored. Again, the real losers with this decision are the students.

Finally, the range of work considered has been limited by downgrading findings from studies that do not incorporate strict random assignment. While we all know that random assignment is the "gold standard" for experimental work, the WWC's over-reliance on this criterion ignores the realities of how school organizations work. With this strict attention to random assignment, other aspects of research designs that are an even greater threat to internal validity can be ignored. The most important of these is no doubt ensuring the fidelity of treatment implementation, making sure that a program is implemented as the developers designed it. The complete rationale behind this concern is too lengthy to include in this letter. I worry that this criterion has involved a sterile and unthinking application of methodological rules that may work for growing corn in a field or worms in a lab. Yet they may be inappropriate, at best, and potentially harmful, at worst, for students in real-life schools and real-life neighborhoods where fidelity of treatment implementation is much more tenuous.

While the concerns noted above have resulted in fewer studies being included, another procedural decision of WWC has resulted in the inclusion of many studies that should probably not have been considered. Unlike most scholarly writing, the WWC material that I have reviewed has surprisingly little relationship to the extant published literature. As I have explored the studies that were reviewed for the website I found that many of them were very difficult to find and, even more shocking, did not appear in peer-reviewed journals, the standard mark of academic respectability. A surprisingly large number of the reports that were considered to have met the evidence standards, either fully or with reservation, were unpublished manuscripts. In some cases, I had to write to the original developers several times before receiving the manuscripts cited in the reviews. In contrast, numerous Direct Instruction articles that were rejected for review have been published in peer-reviewed journals and are available to the general public.

Finally, and perhaps most important, as I have read the studies and their evaluations more closely I have found that *the Clearinghouse's interpretations do not accurately reflect the actual content of the reports*. Below I give some examples of the problems that I have found to date. A very large proportion of the reviews I've read have had little correspondence to the articles themselves.

Quality Problems with the Reviews

One of the programs that received a high rating from WWC in Beginning Reading is *Reading Recovery*®. Most of the other programs reviewed by WWC, such as Direct Instruction, are designed to be used with entire classrooms of students as part of the regular instructional program. In contrast, Reading Recovery (RR) is a "pull-out" program, one that is applied only when students are having trouble in their regular classrooms. The inclusion of the tutorial based Reading Recovery in the Beginning Reading Category is, in fact, rather surprising, given the WWC's stated objective of reporting on "comprehensive … programs," those that enhance "whole school literacy," and "basals/textbooks intended for whole-school/whole-classroom use" (http://ies.ed.gov/ncee/wwc/PDF/BR_protocol.pdf, p. 5).

The WWC website concludes that "*Reading Recovery*® was found to have positive effects on students' alphabetics skills and general reading achievement outcomes and potentially positive effects on comprehension and fluency" (http://ies.ed.gov/ncee/wwc/reports/beginning_reading/reading_recovery/). A careful reading of the articles cited by the website, however, indicates that the reviewers apparently did not understand basic elements of the program or the reported research.

For instance, the article by Baenen and associates (1997) is one of the articles cited as supporting WWC's positive conclusions. In this study first graders were randomly assigned to participate in the pull-out tutoring program or to remain in their regular classroom. Note that the design does not compare Reading Recovery to an alternative curriculum, which would be a rigorous test of the treatment, but instead to a no-treatment situation. Even with this built-in advantage, the authors found remarkably little success, especially in the long-term, for students in the Reading Recovery program. The article concludes (p. 176) that although about one-half of the students in the tutoring program had successfully reached first grade reading levels at the end of the year, success rates declined in subsequent years of implementation. In addition, by third grade, there was no difference in achievement scores of students who had participated in Reading Recovery and other students. Two years after the treatment there were no differences in needs for retention, special education or Chapter 1 assignments. The authors concluded that Reading Recovery was very expensive to implement in relation to the benefits that it provided.

An even more disturbing example involves the article by Iversen and Tunmer (1993), which was also cited by the WWC as supporting the conclusion that RR is effective. The major purpose of this study was to compare the standard RR program to a "modified" program that included explicit instruction in phonological skills. The WWC chose to ignore any results regarding this comparison group "because it was a modified version of the standard program." Unfortunately, this decision resulted in completely ignoring the major conclusions of the article and thus misrepresenting the effectiveness of Reading Recovery and phonologically based programs as well.

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 (with instruction in phonologically based elements added) eventually

caught up with the other children. But the most important finding was that the students in the modified program were able to discontinue the tutoring much earlier. The standard Reading Recovery program was found to be 37 percent less efficient than the modified program. In addition students in the modified program continued to have higher levels of achievement and higher rates of learning at the end of the school year. The authors provide an extensive discussion and additional analysis that demonstrate the fallacy involved in Reading Recovery about the ways in which word recognition skills develop. They clearly conclude that Reading Recovery is not an efficient method for teaching children to read and that phonological training is superior.

Similar, very serious, problems occurred with the review of studies of Direct Instruction, but the mistakes with these articles led the reviewers to dismiss the studies out of hand. One of the most egregious cases involves the study by Carlson and Freeman (2002). The stated reason for dismissing this study was that "there was a confound, with the Direct Instruction intervention being modified or combined with other interventions." In fact, there was only one intervention, and it appears that the reviewer simply did not understand the nature of the intervention well enough to accurately read the article. As with virtually all other studies of Direct Instruction, the analysis, which used sophisticated multilevel techniques, had results strongly in favor of the DI curriculum.

It is possible that the reviewers responded to this section of the description of the intervention, which was termed "RITE" and used the *Reading Mastery* program of Direct Instruction: "In addition to the teaching of skills directly related to the RM curricula, the RITE program also strives to provide teachers with strong classroom management techniques" (p. 143). Strong classroom management is part and parcel of the Direct Instruction approach. Part of the reason that DI is so successful is that it provides not just well designed curricular materials but well developed, research-based guidance on how the curriculum should be administered. These guidelines are well documented in the various guides to the programs and should have been part of the knowledge base of a competent reviewer. The classroom management was not a confounding element of the intervention, but was an integral part of the curriculum.

In other cases, articles regarding a Direct Instruction curriculum were rejected for methodological issues, while those with a different curriculum, but the same methodological approach, were accepted. For instance, Waldron-Soler, et al (2002) found that students instructed in the DI program, *Language for Learning*, had significantly higher achievement than those in programs without explicit language instruction, using analysis of covariance to control for pretest scores. The study was rejected for consideration because "the intervention and comparison groups cannot be considered equivalent at baseline, even with the use of covariates in the analysis." However, the published article includes means and standard deviations for both the experimental and control groups. I calculated simple t-tests and found that in fact there were no significant differences between the groups at pre-test. Such a test was apparently not done by the reviewers, despite their stated conclusion.

To help understand why the Waldron-Soler article was dismissed, I examined studies of Success for All that WWC deemed had met the minimal requirements for inclusion. (Success for All is a reading program that has been found to be effective almost as often as Direct Instruction

and was modeled on many elements of DI.) Three of these papers, none of which had been published in a peer reviewed journal, also used analysis of covariance to adjust for pretest scores (Dianda and Flaherty 1995; Ross, Alberg and McNelis, 1997; and Ross, McNelis, Lewis and Loomis, 1998). Why this adjustment method would pass muster with the unpublished articles regarding Success for All, but be dismissed with the published article regarding Direct Instruction is very unclear and disturbing.

Similarly, two articles by Tobin (2003, 2004) on the *Horizons* DI reading program were rejected because the study groups were supposedly "incomparable." The studies, which were again published in peer reviewed journals, used a technique of matching students in the experimental and treatment groups on pretest measures. Tobin (2003) reported that there were no statistically significant pretest differences between the groups, and, as usual, the results supported the superiority of the DI program. Again I looked at the accepted studies of Success for All to try to understand why Tobin's articles had been rejected. I found that one of the accepted SFA studies (Smith, Ross, Faulks, et al, 1993) also individually matched students on pretest scores. Unfortunately, this unpublished paper only reports the results in graphical form and it was impossible to actually examine the means and standard deviations to test for any differences. Again, I am baffled as to why the unpublished Smith et al piece, which omitted crucial methodological details, would be accepted while the published Tobin pieces, with similar methodology and including the statistical details, were rejected.

Recent innovations to the WWC website have only compounded these errors. The "create your own summary" feature (http://ies.ed.gov/ncee/wwc/Reports/Topic.aspx?tid=01) invites viewers to create a summary of the reports by simply clicking a button. The resulting page lists interventions in the order of the "improvement index" score, implying that those at the top of the list are the most effective while those at the bottom of the list are least effective.

One of the programs often listed near the top of the Beginning Reading lists is Reading Recovery; but, as explained above, the reviews of the program were fatally flawed and inaccurate. Another program rated near the top is Early Intervention in Reading (EIR). Like Reading Recovery, EIR involves special tutoring for low performers, yet, as with the reviews of Reading Recovery, the judgment of EIR's effectiveness is questionable. The conclusion regarding EIR is based on one unpublished study, involving 59 students. Half of these students received tutoring in addition to their classroom instruction and their reading achievement was compared to those who received no additional systematic instruction. While the children who received tutoring did better than those who did not, there is no way to tell if EIR is superior to another type of intervention or if it is simply the additional practice time or attention that made the difference.

In short, given the flawed process of selecting studies for review and the extremely poor quality of the review process, any "effectiveness ratings" are undoubtedly inaccurate and misleading. To prominently display these ratings and encourage their use does a great disservice to schools and parents throughout the country and will eventually, I believe, harm Mathematica's reputation of scientific integrity.

What Should Be Done

I bring these very serious concerns to your attention in the hope that you will correct the mistakes that have been made and establish procedures that will prevent them from occurring in the future. The articles by Slavin and McArthur provide a number of very good suggestions. For instance, Slavin provides very cogent suggestions regarding the development of program ratings that more accurately reflect the strength of the evidence, altering the decision process regarding acceptable study designs, and giving greater weight to larger studies and those of longer duration. McArthur suggests that reviews be limited to studies that have passed peer review, that measures be adjusted to more accurately reflect the concepts that are supposedly being rated, that the method of rating programs more accurately reflect the results and, most important, that reviewers be familiar with the subject area and that all summaries be reviewed by independent experts in the area before publication on the website.

I fully support these suggestions and applaud their common notion that the WWC should use standard methods of scientific procedure and review and that those who do the reviews should be familiar with the areas under study. Acknowledging that there is overlap with the suggestions of Slavin and McArthur, I will list my own recommendations:

1) Make the criteria used in searching for studies explicit and replicable.

2) Concentrate on peer reviewed studies first.

3) Ensure that all studies accepted are available in full text for review by others.

4) Use all the available evidence regarding interventions, with no arbitrary cut-off dates or age or grade ranges that unfairly exclude a body of evidence.

5) Ensure that reviewers are knowledgeable in the substantive areas that they are reviewing as well as in the methodological details.

6) Develop standard quality control mechanisms, such as two reviewers for every article with a third if results differ and a peer review process before publication of ratings.

7) Develop procedures to ensure comparability from one area of the review to another.

8) Compare results that are obtained with the already well developed body of meta-analyses and literature reviews. When WWC's results differ from these reviews, investigate why and adjust results as needed.

9) Use measures of effect size that are more commonly used within the social sciences and comparable to those used in other review studies. Include elements related to sample size, statistical significance, substantive significance, and length of intervention.

10) Weigh issues of both internal and external validity in decisions regarding acceptable designs and recognize the wide variety of approaches that can and must be used within real-life educational settings.

11) Organize topics in ways that reflect how schools actually operate, such as core reading programs, supplemental programs, etc.

While all of these changes are vital to ensuring that future reviews are accurate, they do not deal with the faulty reviews that are already posted on the website. Given the very serious concerns that Slavin, McArthur, and I (and perhaps others of whom I'm not aware) have raised regarding the content of the WWC website, I believe that the reviews and ratings that are now posted must be removed until they can be reviewed for accuracy with well designed quality control measures. The problems with the material posted on the website appear to be much more widespread than can be handled in a piece-meal fashion. To allow inaccurate material to remain does a disservice to schools, families and students. I know that the vast majority of the work

posted on the website was not done under Mathematica's guidance. Continuing to post such faulty material can only, unfairly, sully Mathematica's reputation.

Thank you for attending to my concerns. Please know that they are presented in the spirit of the intent behind the What Works Clearinghouse – a desire to provide schools and families with the most accurate information. If you wish to discuss any of my concerns further, please do not hesitate to contact me. I look forward to hearing from you.

Sincerely,

Jean Stockard, Ph.D.
Director of Research
jstockard@nifdi.org
Toll free phone # 877-485-1973

6) Follow-up letter, JS to Mathematica, July 31, 2008, sent with return receipt via Fed-ex.

July 31, 2008

Anita A. Summers, Chairperson and
Paul T. Decker, President and CEO
Mathematica, Policy Research Inc.
P.O. Box 2393
Princeton, NJ 08543-2393

Dear Drs. Summers and Decker:

It has now been more than a month since I wrote you regarding my concerns with the material posted on the What Works Clearinghouse website. However, I have received no response or even an acknowledgement of my letter. In case you did not receive the original communication, I have included a copy in this package. I have also sent this letter via Federal Express so that I may confirm that it has been received.

As you know, I believe that the analyses of curricular material posted on the WWC website are fatally flawed. They provide very erroneous information to educators and parents. Because Mathematica only recently assumed responsibility for the WWC and because I know that Mathematica has had, in the past, a reputation for high quality work, I felt it important to let you know of these problems privately so that you could address them.

The lack of response to my earlier letter suggests, however, that my concerns, as well as those of others, have not been given serious credence. Given the serious nature of the issues involved, unless I hear from you within the next 2 weeks, I will have to pursue alternative means of dealing with this matter. I am available to speak via phone (toll-free at 877-485-1973) or to communicate via e-mail (jstockard@nifdi.org) if you prefer those means of communication. I hope to hear from you soon.

Sincerely,

Jean Stockard, Ph.D.
Director of Research
jstockard@nifdi.org
Toll free phone # 877-485-1973

7) Decker to JS, August 6, 2008

Response to Your Letter

Dear Dr. Stockard:

Thank you for prompting me regarding a reponse to your concerns about the What Works Clearinghouse (WWC). I received your earlier letter and reviewed it carefully. As you know, your original letter contained a substantial amount of material as well as 11 specific recommendations to change the criteria and processes supported by the Clearinghouse. I therefore have asked for input from both our team working on the project as well as our client for this work, the Institute of Education Sciences (IES) at the U.S. Department of Education, to develop an appropriate response to your letter.

As you probably know already, Mathematica is administering the WWC under contract to IES, and our work is bound by the statement of work contained in the contract. Furthermore, we work closely with IES to determine how best to fulfill the statement of work as the details of the work unfold. Hence, any potential changes that Mathematica could make in response to your recommendations, including changes in the WWC web site, may be limited by the contract or by the guidance of our client.

Thank you for your attention to the What Works Clearinghouse initiative. I appreciate your thoughtful review, and I look forward to sending you a lengthier response.

Paul Decker

_____

Paul Decker
President and Chief Executive Officer
Mathematica, Inc.
Box 2393
Princeton, NJ 08543
609-275-2290

8) JS to Decker, August 7, 2008

Dear Dr. Decker:

Thank you for your response to my letters. I look forward to receiving your lengthier reply. Might you have any idea of the time frame for this response?

Again, thank you for attending to my concerns.

Sincerely,

Jean Stockard

9) A.S. to J.S., September 6, 2008

Ms. Stockard: Thank you for copying me on your Aug. 29 letter to Paul Decker. I am
following up on it with him, and am assured that he is addressing the issues meticulously. I will continue to connect with him on this matter.

Sincerely,

Anita A. Summers

10) Mathematica reply, September 8, 2008

Note that this has two part: The first is a cover letter from Paul Decker the President and CEO of Mathematica. The second is a more detailed response to the June 25 letter, apparently prepared by Mark Dynarski. This is referred to in the text as Dynarski, 2008.

MATHEMATICA
Policy Research, Inc.

Paul T. Decker
President and Chief Executive Officer

P.O. Box 2393
Princeton, NJ 08543-2393
Telephone (609) 799-3535
Fax (609) 799-0005
www.mathematica-mpr.com

September 8, 2008

Dr. Jean Stockard, Ph.D.
Director of Research
National Institute of Direct Instruction
P.O. Box 11248
Eugene, OR 97440

Dear Dr. Stockard:

Thank you for your June 25 letter concerning the What Works Clearinghouse (WWC). Mathematica Policy Research wants to ensure the quality and accuracy of all information contained in the WWC, and it works with study authors, curriculum developers, and consumers of the WWC to correct inaccurate information. We therefore appreciate your interest in and feedback on the WWC.

The WWC was designed to establish and apply a set of standards to identify rigorous research related to the impacts of education interventions. Studies that meet these standards are considered to have designs with causal validity, which gives practitioners increased confidence that the results of the study actually reflect the true impact of the intervention being examined. These standards were developed by leading experts in education research methodology, and they are applied through a systematic review process that includes repeated checks for quality and accuracy. They represent research criteria that we believe to be vital to the purpose of the WWC.

I have closely reviewed the concerns in your letter. I have also asked for input from Mark Dynarski, the director of the WWC effort, and representatives from the Institute for Education Sciences at the U.S. Department of Education. Based on my review and the input I received, I am convinced that the standards being used by the WWC adequately reflect the original mission of the WWC, as outlined above, and that those standards are being applied appropriately and consistently. In the attached document, Mark Dynarski provides more detailed responses to the issues you raise concerning the WWC's review of Direct Instruction.

Your letter makes several suggestions for the WWC moving forward. Some suggestions are counter to the systematic review process that is the foundation of the WWC. For example, as explained in the attachment, the criteria for date and age ranges being reviewed are established with specific rationales, and the systematic review process gives equal weight to peer-reviewed and non-

peer-reviewed sources. As we expand the WWC's activities into new topic areas, we will be mindful of the suggestions you raised in your letter.

I hope this information clarifies the WWC mission and the way in which the WWC operations support that mission.

Sincerely,

Paul

Paul T. Decker

RESPONSES TO CONCERNS IN 6/25/08 LETTER FROM JEAN STOCKARD

Contrasting conclusions between WWC and extant literature

The letter states that WWC conclusions differ from extant literature including meta-analyses and literature reviews. It is important to note that the WWC is designed to produce a systematic review of literature. A sound definition of a "systematic review" is in the recent publication *Knowing What Works in Health* Care by the Institute of Medicine (IOM):[1] *A systematic review is a scientific investigation that focuses on a specific question and uses explicit, preplanned scientific methods to identify, select, assess, and summarize similar but separate studies.* (pg. 82)

Consistent with the recommendations in the IOM report, the WWC applies evidence-based methodological standards consistently to each study it reviews. These standards, which are available on the WWC website, were developed by leading education research methodologists.

The systematic review conducted by the WWC goes beyond the procedures performed in meta-analyses. Again using the IOM definition, a meta-analysis "quantitatively combines the results of similar studies in an attempt to allow inference from the sample of studies included to the population of interest" (pg. 82). However, as Robert Slavin has noted, meta-analyses rarely describe even one study in any detail.[2] The WWC uses meta-analytic techniques to summarize the results of studies that meet WWC standards. However, the important distinction is that the WWC uses a rigorous set of standards, applied consistently, to determine *which* studies are included in the meta-analytic computations. This ensures that WWC summary measures are based only on studies with causal validity.

It is not surprising that WWC findings may differ from those of other analyses. Some meta-analyses may not have such rigorous standards for including studies; others may have standards that differ from those used by the WWC. Because of these differences, the WWC cannot judge the results of its systematic reviews by how they compare to other analyses. Rather, it works to ensure that its standards appropriately identify studies with strong causal validity and applies those standards consistently to each study reviewed.

---

[1] Institute of Medicine (IOM). *Knowing What Works in Health Care: A Roadmap for the Nation.* Washington, DC: The National Academies Press. 2008.

[2] Slavin, R.E. (1995). Best evidence synthesis: an intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology*, Vol.48, No.1 pp.9-18.

Inclusion and exclusion procedures

Limiting studies to 1985 or later

The WWC's default time period for reviews is a study publication date of 1985 or later. This timeframe, which was established in 2005, is used for two reasons. First, by limiting reviews to research to this time period, WWC reviews reflect reasonably current research. In particular, the time period range ensures that effect sizes and improvement indices are based on a counterfactual condition that reflects classrooms as they operate within a recent time period. Second, the timeframe ensures that the research reviewed is examining versions of interventions that are most likely to be available to practitioners today.

WWC principal investigators have the option to expand the period for which studies can be reviewed, if they believe that important research will be excluded. The principal investigators for the Beginning Reading area chose to maintain the default period in large part to maintain currency with the classroom context for beginning readers. For instance, the fact that preschool enrollment has increased,[1] combined with the fact that more preschool and kindergarten programs run full-day,[2] means that students in the early grades may be better prepared to receive reading instruction today than students 25 years ago. Moreover, it is possible that any changes in reading readiness over this period have not been evenly distributed, since differences in reading ability by socioeconomic status and race are apparent at the kindergarten level.[3] Other contextual factors have changed over the past 20 years, including advances in teacher training, increases in home literacy activities, and changes in the content of and variety of curricula used in classrooms.

Any of these changes could have implications for the effectiveness of an intervention. If school readiness has increased, than an intervention that was effective 25 years ago may not be effective in more recent years. If teachers are receiving stronger training and using newer curricula, the counterfactual condition against which interventions were measured 25 years ago have changed, and possibly with it the magnitude of its effects. The Beginning Reading principal investigators judged that they had an inadequate basis for

---

[1] The proportion of 3 and 4 year olds enrolled in school increased from 37 percent in 1980 to 56 percent in 2006. See Snyder, T.D., Dillow, S.A., and Hoffman, C.M. *Digest of Education Statistics 2007*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, 2008.

[2] The proportion of pre-kindergartners and kindergartners attending full-day pre-kindergartners or kindergarten increased from 32 percent in 1980 to 59 percent in 2006. See Snyder, Thomas D. *Mini-Digest of Education Statistics, 2007*. Washington, DC: National Center for Education Statistics, Institute of Educational Sciences, U.S. Department of Education, 2008.

[3] For example, the Early Childhood Longitudinal Study of Kindergarteners (ECLS-K) found gaps in the reading knowledge and skills of kindergarteners by race: black and Hispanic children scored just under half of a standard deviation below whites on a test of reading knowledge and skills. Analysis of the ECLS-K and other surveys (i.e., Children of the National Longitudinal Study of Youth and the Infant Health and Development Program) show that socioeconomic status accounts for about half of the standard deviation of racial differences in reading test scores. See Duncan, G. and K. Magnuson. "Can Family Socioeconomic Resources Account for Racial and Ethnic Test Score Gaps?" *Future of Children*, Vol. 15, No. 1, Spring 2005.

assuming that effects of interventions measured more than 20 years ago would be experienced if schools adopted those interventions today.

Restrictive grade range for Beginning Reading studies

The letter raised concerns that the WWC's exclusion of research conducted on children outside of the kindergarten to 3rd grade age range was too restrictive. Because the reviews are focused on assessing interventions for beginning reading, the principal investigators, in coordination with the Institute of Education Sciences (IES), concluded that the review should focus on intervention effects for children in kindergarten through third grade (roughly ages 5 to 8). Some studies examine effects of interventions for students within the specified grade range and also students in higher grades. These studies are included in the review when the WWC is able to isolate the effects for the students who fall within the Beginning Reading grade range. The studies are excluded from the review if the results cannot be disaggregated to isolate effects for relevant grade range.

The grade range criterion is important for the integrity of the review process. The WWC-computed improvement indices and effect sizes are intended to reflect the effect of the intervention on the population in question. Including students above the topic area age range could lead to misstatements about intervention effects on children within the grade range.

It should be noted that the WWC attempts to determine a study's effects for the relevant grade range. When authors present findings aggregated across a broader age range and indicate that findings were analyzed for the relevant age range, it is standard WWC procedure to contact authors to request findings disaggregated for the grade range. Sometimes design limitations or other factors preclude authors from providing disaggregated results. In such cases, the WWC excludes the study.

Fidelity of treatment implementation

The letter notes that the WWC review process may downplay implementation fidelity. Definitions of implementation fidelity vary and many studies include little information to gauge fidelity, especially information about whether an intervention has been implemented within normal operating regimes of districts, schools, and teachers, not under specialized laboratory conditions. Moreover, there is no standard metric with which to rate and assess fidelity across studies that assures comparability.

The WWC's approach emphasizes the importance of replicated findings, which ensures that any one study in which fidelity issues may have arisen are averaged with findings from other studies. Intervention reports include an "extent of evidence" classification that allows practitioners to place more weight if they choose on interventions for which the extent of evidence is large, meaning the results are drawn from multiple studies and a large number of classrooms and students.

Concerns about interventions and studies reviewed by the WWC

Reading Recovery

The letter expressed concern that Reading Recovery is an intervention outside of the Beginning Reading protocol. The Beginning Reading protocol states that interventions that target specific populations (for example, readers below grade level, and at-risk students) are eligible for the review. Reading Recovery is a short-term tutoring intervention program intended to serve the lowest achieving first-grade students (i.e., those in the bottom 20 percent). As such, it falls within the Beginning Reading protocol.

The letter expressed concerns that the reviews of studies of Reading Recovery mischaracterized the findings from those studies. For both the Baenen et al. (1997) and Iversen and Tunmer (1993) studies, the results presented by the WWC review represent the findings when the WWC standards and procedures are applied to these studies.

With respect to the Baenen et al. study, it is important to note that the beginning reading protocol prioritized one-year results. In effect, the Beginning Reading review is only intended to examine whether beginning reading interventions have an effect within one year. This one-year period is applied consistently to each study reviewed to ensure the results can be compared across studies and interventions. The Baenen et al. study's two- and three-year general reading achievement measures are not ignored, however; they are presented in Appendix A4.4 of the Reading Recovery Technical Appendices (http://ies.ed.gov/ncee/wwc/pdf/techappendix01_209.pdf).

With respect to the Iversen and Tunmer study, and consistent with the protocol, the WWC examined the results most relevant to the question of whether Reading Recovery improves reading proficiency compared to a reasonable counterfactual. That the study examined other comparisons is not ignored however. Appendices A4.1, A4.2 and A4.3 present results from other comparison groups. As with any study it reviews, the WWC does not base the findings of its review on the conclusions drawn by the authors.

We disagree that studies that compare an intervention to a no-treatment condition (as was done in Baenen et al.) provides a "built-in advantage," as the letter suggests. Many practitioners are interested in knowing whether an intervention is effective relative to customary classroom practices. The WWC reviews studies comparing treatments to no-treatment as well as studies comparing one treatment to another. In each case, the counterfactual is clearly documented in the review.

Exclusion of Reading Mastery Program

The letter expresses concern that the Reading Mastery Program is excluded from review by the WWC. The WWC has reviewed studies of Reading Mastery. A report summarizing the results of those reviews was published on the WWC website on August 12, 2008.

Inclusion of unpublished manuscripts for review

The letter suggests that by reviewing unpublished manuscripts, the WWC reviews studies of lower quality. Reviewing only studies that have passed a peer reviewing process could neglect unpublished information that may contain important and different findings. Mark Lipsey and David Wilson caution against this type of publication bias in *Practical Meta-Analysis* (2001). The damage such publication bias can cause in a systematic review is substantial. In *Knowing What Works in Healthcare*, the IOM notes that publication bias is well-established and that systematic reviews need to take steps to counter it, because otherwise "harmful interventions may appear to be worthwhile and beneficial interventions may appear to be useless" (page 97). For this reason WWC procedures are explicitly designed to review evidence from sources other than those published in journals. However, regardless of publication status, studies must meet the same evidence standards.

The letter asserts that the WWC should make available to the public all the research it reviewed. The more important issue is that the WWC provides an explicit citation to the study and thereby enables readers to obtain studies they wish to review. Unpublished information obtained from authors in response to queries that arise in reviews also are made available to requestors, and authors sign a form indicating that the information they provide to the WWC is available to the public.

Exclusion of Carlson and Francis (2002) study

The letter expressed concern over the WWC review's conclusion that there was a confound with the Direct Instruction intervention in Carlson and Francis (2002).[1] The WWC standard regarding intervention confounds was established to ensure that the results in a WWC review reflect what educators can expect if they implement the intervention being reviewed. A careful reading of Carlson and Francis indicates that findings cannot be separated into effects of Reading Mastery alone and effects of Reading Mastery supplemented by the support provided to teachers through the RITE program.

Exclusion of Waldron-Soler et al. (2002) study

As described in the protocols for the WWC Early Childhood Education and Beginning Reading reviews, to establish baseline equivalence, treatment and comparison groups must differ on the pretest measures by less than half a standard deviation, or the differences must be insignificant in an adequately powered statistical test. The Waldron-Soler (2002) study reported pretest differences exceeding half a standard deviation on at least two measures.

---

[1] The letter references Carlson and Freeman (2002). We have no record of a Carlson and Freeman (2002) study, and have assumed the letter refers to 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 Direct Instruction*, 3(1), 29-50.

The same standards were applied to the three Success for All studies you cite (Dianda and Flaherty, 1995; Ross, Alberg and McNelis, 1997; and Ross, McNelis, Lewis and Loomis, 1998). Unlike the Waldron-Soler study, pretest differences between the treatment and comparison groups were less than one half of a standard deviation.

Exclusion of Tobin (2003, 2004) studies

The Tobin (2003) study indicated that treatment and comparison groups were substantively different on pretest measures, with the differences exceeding half a standard deviation. Because the Tobin (2004) study follows the same students for additional time, it did not meet standards for the same reason.

The same standards were applied to the Success for All study cited in the letter (Smith, Ross, Faulks et al., 1993). Unlike the Tobin studies, pretest differences between the treatment and comparison groups were not statistically significantly different, and they were less than one half of a standard deviation. In keeping with WWC procedures, the WWC obtained information on pretest differences through communication with the study authors.

The "Create your own summary" feature

The letter expresses concern about the "create your own summary" feature of the WWC website because the results sort interventions by the magnitude of the improvement index. Because the improvement index calculations accurately reflect the application of the WWC evidence standards and effect size computations to the studies of those interventions, the WWC believes it is informative and useful to practitioners to sort results by improvement index.

The improvement index is one summary measure of the effect of beginning reading interventions. Users can also sort interventions alphabetically (by intervention name), by evidence rating, and by extent of evidence indices.

Appendix B
A Partial List of Studies of *Reading Mastery* and its Precursors Completed before 1985

Note: The letters in parentheses at the end of a citation refer to the source from which it was obtained. These sources are listed at the end of this appendix.

Apffel, J.A., J. Kelleher, M.S. Lilly, and R. Richardson. 1980. Developmental reading for moderately retarded children. *Education and Training of the Mentally Retarded* 10: 229-235. (b) (g)

Beck, I. L. and E. S. McCaslin. 1978. *An analysis of dimensions that affect the development of code-breaking ability in eight beginning reading programs*. LRDC Report No. 1978/6. Pittsburgh: University of Pittsburgh Learning Research and Development Center. (c)

Becker, W. C. and D. W. Carnine. 1980. Direct instruction: An effective approach to educational intervention with the disadvantaged and low performers. Pp. 429-473 in B.B. Lahey and A. E. Kazdin (eds.), *Advances in Clinical Child Psychology, Volume 3*, New York: Plenum. (a)

Becker, W. C. and S. Engelmann. 1976. *Analysis of achievement data on six cohorts of low-income children from 20 school districts in the University of Oregon Direct Instruction Follow Through Model.* Eric Document Reproduction Service No. ED 145922. (a)

Bock, G. and L. B. Stebbins. 1977. *Education as experimentation: A planned variation model, Volume IV-B effects of Follow Through models*. Cambridge, MA.: Abt Associates (ERIC Document Reproduction Service No. ED 148491). (a)

Booth, A., D. Hewitt, W. Jenkins, and A. Maggs. 1979. Making retarded children literate: A five-year study. *The Australian Journal of Mental Retardation* 5: 257-260. (d) (i)

Bowers, W. M. l972. *An evaluation of a pilot program in reading for culturally disadvantaged first grade students*. Doctoral dissertation, University of Tulsa, 1972. (Eric Document Reproduction Service No. ED 073439). (a) (b)

Bracey, S., A. Maggs, and P. Morath. 1975. Effects of a direct phonic apporach in teaching reading with six moderately retarded children: Acquisition and mastery learning stages. *Slow Learning Child* 22: 83-90. (b) (d) (i)

Branwhite, A. B. 1983. Boosting reading skills by Direct Instruction. *British Journal of Educational Psychology* 53: 291-298. (b) (c) (d) (g) (i)

Carnine, D. 1977. Phonics versus look-say: Transfer to new words. *The Reading Teacher* 30: 636-640. (h)

Carnine, D. 1980. Phonic versus whole-word correction procedures following phonic instruction. *Education and Treatment of Children* 3: 323-329. (h)

Carnine, D. and R. Gersten. 1983. *Effectiveness of Direct Instruction in teaching selected reading comprehension skills: Preliminary Draft*. Paper presented at American Educational Research Association (April 11-15, 1983). (b)

DuPree, T. J. 1976. Brief history of Cherokee schools, 1804-1976. *BIA Education Research Bulletin* 4: 3-11. (ERIC Document Retrieval Service No. ED 127051). (a)

Gersten, R., W.C. Becker, T. J. Heiry, and W. A. T. White. 1984. Entry IQ and yearly academic growth of children in Direct Instruction programs: A longitudinal study of low SES children. *Educational Evaluation and Policy Analysis* 6: 109-121. (e)

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

Gersten, R., A. Brockway, and N. Henares. 1983. The Monterey DI program for students with limited English (ESL). *Direct Instruction News*, 2(4): 8-9. (a)

Gordan, M.B. (ed.) 1971. *DISTAR instructional system: Summaries of case studies on the effectiveness of the DISTAR instructional system*. Chicago: Science Research Associates. (a)

Gregory, R. P. and B. G. Warburton (1983). *DISTAR Reading* and remedial children in an infant school. School Psychology International 4: 169-172. (i)

Haring, N.G. and D. A. Krug. 1975. Evaluation of a program of systematic instructional procedures for extremely poor retarded children. *American Journal of Mental Deficiency* 79: 627-631. (b) (i)

Kastner, S. and M. Hollingshead. 1973. *An evaluation of ESEA Title I Programs, Community School District 15.* New York: New York University, Center for Educational Research and Field Services, School of Education. (ERIC Document Reproduction Service No. ED 087842). (a)

Kaufman, M. 1973. *The effect of the DISTAR Instructional System: An evaluation of the 1972-1973 Title I program of Winthrop, Massachusetts*. (ERIC Document Reproduction Service No. ED 110171). (a)

Kaufman, M. 1974. *The effect of the DISTAR Instructional System: An evaluation of the 1973-1974 Title I program of Winthrop, Massachusetts*. (ERIC Document Reproduction Service No. ED 110170). (a)

Lewis, A. 1982. An experimental evaluation of direct instruction programme with remedial readers in a comprehensive school. *Educational Psychology* 2: 121-135. (g)

Lloyd, J., D. Cullinan, E.D. Heins, and M.H. Epstein. 1980. Direct instruction: Effects on oral and written language comprehension. *Learning Disabilities Quarterly* 3: 70-76. (g)

McCabe, T.A. 1974. *The DISTAR Reading and Language Program: Study of its effectiveness as a method for the initial teaching of reading*. Doctoral dissertation, University of Massachusetts, 1974. (ERIC Document Retrieval Service No. ED 102498).

Meyer, L. A. 1983. *Long-term academic effects of Direct Instruction Follow Through. Technical report No. 299.* Champaign: University of Illinois at Urbana-Champaign, Center for the Study of Reading. (ERIC Document Retrieval Service No. 237932). (a)

Newark (NJ) Board of Education. 1974. *Program to improve the informational processoing of children with reading and learning problems*. ERIC 106 826. (b)

Ogletree, E. J. 1976. *A comparative study of the effectiveness of DISTAR and eclectic reading methods for inner-city children*. Chicago: Chicago State University. (ERIC Document Reproduction Service No. ED 146544). (a) (b)

Ogletree, E.J. 1977. *Does DISTAR meet the reading needs of inner-city kindergarten pupils?* ERIC 146 303. (b)

Rawl, R. K. and F. S. O'Tuel. 1982. A comparison of three prereading approaches for kindergarten students. *Reading Improvement* 19: 205-211. (a) (b)

Richardson, E., B. Dibenedetto, A. Christ, M. Press, and B. Winsbert. 1978. An assessment of two methods for remediating reading deficiencies. *Reading Improvement* 15: 82-95. (a) (b) (d) (g) (i)

Sassenrath, J. M. and R. E. Maddux. 1974. Language instruction, background, and development of disadvantaged kindergarten children. *California Journal of Educational Research* 25: 61-68. (a)

Serwer, B. L., B. J. Shapior, and P. P. Shapiro. 1973. Comparative effectiveness of four methods of instruction on the achievement of children with specific learning disabilities. *Journal of Special Education* 7: 241-249. (b) (d)

Singer, B. 1973. *The effects of structured instruction on kindergarten pupils: Final report.* Washington, D.C.: U.S. Office of Education. (ERIC Document Reproduction Service No. ED 087564). (a)

Stebbins, L. B., R.G. Pierre, E. C. Proper, R. B. Anderson, and T. R. Cerva. 1976. *Education as experimentation: A planned variation model* (Vols. 3A-3B). Cambridge, MA: Abt Associates. (ERIC Document Reproduction Service No. ED 148489). (a)

Stein, C. L. and J. Goldman. 1980. Beginning reading instruction for children with minimal brain dysfunction. *Journal of Learning Disabilities* 13: 52-55. (b) (g) (i)

Summerell, S. and G. G. Brannigan. 1977. Comparison of reading programs for children with low levels of reading readiness. *Perceptual and Motor Skills* 44: 743-746. (a) (b) (d) (g)

Williamson, F. 1970. *DISTAR reading: Research and experiment*. (ERIC Document Reproduction Service No. ED 045318). (a) (b)

Note: Letters in parentheses refer to the source(s) providing the citation.

a. Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2003. Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research* 73: 125-230.
b. Adams, Gary L. and Siegfried Engelmann. 1996. *Research on Direct Instruction: 25 years beyond DISTAR*. Seattle: Educational Achievement Systems.
c. McGraw-Hill, 2002. "Appendix: Research supporting instruction in *Reading Mastery*: A Selected Annotated Bibliography." *Results with Reading Mastery*. ERIC ED 469 523.
d. Schieffer, Cheryl, Nancy E. Marchand-Martella, Ronald C. Martella, Flint L. Simonsen, and Kathleen M. Waldron-Soler. 2002. An analysis of the *reading mastery* program: Effective components and research review. *Journal of Direct Instruction* 2: 87-199.
e. Comprehensive School Reform Quality Center. 2006. *CSRQ Center Report on Elementary School Comprehensive School Reform Models*. Washington, D.C. American Institutes for Research.
f. NIFDI research staff found this citation independently.
g. Grossen, Bonnie. n.d. "References for *Reading Mastery* Reading Review," The Research Base for *Reading Mastery*, SRA, Eugene, Oregon: University of Oregon. (http://darkwing.uoregon.edu/~adiep/rmref.htm).
h. National Reading Panel. 2000. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washing ton, D.C.: National Institute of Child Health and Human Development.
i. Kinder, D., R. Kubina, N.E. Marchand-Martella. 2005. Special education and Direct Instruction: An effective combination. *Journal of Direct Instruction* 5: 1-36.

Appendix C
Studies of *Reading Mastery* Published 1985 or Later But Not Included in the WWC Review

Note: The letters in parentheses at the end of a citation refer to the source from which it was obtained. These sources are listed at the end of this appendix.

Ball, E. and B. Blachman. 1991. Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly* 26: 49-66. (d)

Bateman, B. 1991. Teaching word recognition to slow-learning children. *Reading, Writing, and Learning Disabilities* 7: 1-16. (b)

Brent, G. and N. DiObilda. 1993. Effects of curriculum alignment versus direct instruction on urban children. *Journal of Educational Research* 86: 333-338. (a)

Brett, A., L. Rothlein, and M. Hurley. 1996. Vocabulary acquisition from listening to stories and explanation of target words. *Elementary School Journal* 96: 415-422. (d)

Bryne, B. and R. Fielding-Barnsley. 1991. Evaluation of a program to teach phonemic awareness to young children. *Journal of Educational Psychology* 83: 451-455. (d)

Butler, P.A. 2003. Achievement outcomes in Baltimore City Schools. *Journal of Education for Students Placed at Risk* 8: 33-60. (e)

Chamberlain, L.A. 1987. Using DI in a Victoria, B.C. resource room. *ADI News* 7 (1): 7-8. (i)

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

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

Cooke, N.L., S.L. Gibbs, M. L. Campbell, and S. L. Shalvis. 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: 139-151. (i)

Cunningham, A. 1990. Explicit versus implicit instruction in phonemic awareness. *Journal of Experimental Child Psychology* 50 :249-444. (d)

Dale, P.S. and K.N. Cole. 1988. Comparison of academic and cognitive programs for young handicapped children. *Exceptional Children* 54: 439-447. (i)

Darch, C. and E. Kameenui. 1987. Teaching critical reading skills to learning disabled children. *Learning Disability Quarterly* 10: 82-92. (g)

Diobilda, N. and G. Brent. 1986. Direct Instruction in an urban school system. *Reading Instruction Journal* 29: 2-5. (e)

Dowdell, T. 1996. *The effectiveness of Direct Instruction on the reading achievement of sixth graders*. (ERIC Document Retrieval Service No. ED 396268). (a) (d)

Fredrick, L.D., M.C. Keel, and J. H. Neel. 2002. Making the most of instructional time: Teaching reading at an accelerated rate to students at risk. *Journal of Direct Instruction* 2: 57-63. (f)

Gersten, R. and D. Carnine. 1986. Direct Instruction in reading comprehension. *Educational Leadership* 44: 68-78. (b)

Gersten, R., C. Darch, and M. Gleason. 1988. Effectiveness of a Direct Instruction academic kindergarten for low-income students. *The Elementary School Journal* 89: 227-240. (e) (h)

Gersten, R., T. Keating, and W. Becker. 1988. The continued impact of the direct instruction model: Longitudinal studies of follow through students. *Education and Treatment of Children* 11: 318-327. (h)

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

Haskell, D. W., B. R. Foorman, and P. R. Swank. 1992. Effects of three orthographic/phonological units on first-grade reading. *Remedial and Special Education* 13: 40-49. (c) (d)

Herrera, J. A., C. H. Logan, P. G. Cooker, D. P. Morris, and D. E. Lyman. 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: 71-89. (d)

Jenkins, J. R., B. Matlock, and T. A. Slocum. 1989. Two approaches to vocabulary instruction: The teaching of individual word meanings and practice in deriving word meaning from context. *Reading Research Quarterly* 24: 215-235. (d)

Juel, C. 1988. Learning to read and write: A longitudinal study of 54 children from first through fourth grade. *Journal of Educational Psychology* 80: 437-447. (d)

Juel, C. and C. Minden-Cupp. 2000. Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly* 35: 458-492. (d)

Kamps, D.M., H.P. Wills, C.R. Greenwood, S. Thorne, J.F. Lazo, J.L. Crockett, J.M. Akers, and B. OL. Swaggart. 2003. Curriculum influences on growth in early reading fluency for students with academic and behavioral risk. *Journal of Emotional and Behavioral Disorders* 11: 211-224. (f)

Kaiser, S. K. Palumbo, R.C. Bialozor, and T. F. McLaughlin. 1989. Effects of Direct Instruction with rural remedial students: A brief report. *Reading Improvement* 26: 88-93. (b) (c) (g)

Keel, M.C., L. D. Frederick, T. A. Hughes, and S.H. Owens. 1999. Using paraprofessionals to deliver Direct Instruction reading programs. *Effective School Practices* 18: 16-22. (e)

Kuder, S.J. 1990. Effectiveness of the DISTAR reading program for children with learning disabilities. *Journal of Learning Disabilities* 23: 69-71. (b) (d) (i)

Kuder, S. J. 1991. Language abilities and progress in a Direct Instruction reading program for students with learning disabilities. *Journal of Learning Disabilities* 24: 124-127. (b) (d)

Leach, D. and S. Siddall. 1990. Parental involvement in the teaching of reading: A comparison of hearing reading, paired reading, pause, prompt, praise, and direct instruction methods. *British Journal of Educational Psychology* 60: 349-355. (h)

Ligas, M. R. and D. W. Vaughan. 1999. *Alliance of quality schools: 1998-99 evaluation report.* Broward , Florida: Broward County Schools. (a)

MacIver, M.A., E. Kedmper, and S. Stringfield. 2000. *The Baltimore Curriculum Project: Fourth year report. Report for the Abell Foundation*. Baltimore: Johns Hopkins University, Center for Social Organization of Schools. (a)

Marston, D., S. Deno, D. Kim, K. Diment, and D. Rogers. 1995. Comparison of reading intervention approaches for students with mild disabilities. *Exceptional Children* 62: 20-37. (h) (i)

Mathes, P.G. and T.J. Proctor. 1988. Direct Instruction for teaching "hard to teach" students. *Reading Improvement* 25: 92-97. (b)

Neely, M. 1995. The multiple effects of whole language, precision teaching, and Direct Instruction on first-grade story-reading. *Effective School Practices* 14: 33-42. (d)

O'Connor, R. E. and J. R. Jenkins. 1995. Improving the generalization of sound/symbol knowledge: Teaching spelling to kindergarten children with disabilities. *The Journal of Special Education* 29: 255-275. (i)

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

Rawl, R. K. and F. O'Tuel. 1982. A comparison of three prereading approaches for kindergarten students. *Reading Improvement* 19: 205-211. (d)

Reutzel, D. R. and P. M. Hollingsworthy. 1993. Effects of fluency training on second graders' reading comprehension. *Journal of Educational Research* 86: 325-331. (d)

Ross, S., J. Nunnery, E. Goldfeder, A. J. McDonald, R. Rachor, M. Hornbeck, et al. 2004. Using school reform models to improve reading achievement: A longitudinal study of Direct Instruction and Success for All in an urban district. *Journal of Education for Students Placed at Risk* 9: 357-389. (e)

Salerno, C. 1992. A comparison of classrooms using a meaning-centered approach and a code-centered approach. *ADI News* 11 (2): 26-29. (d)

Sexton, C. W. 1989. Effectiveness of the DISTAR Reading I Program in developing first graders' language skills. *Journal of Educational Research* 82: 289-293. (a) (b) (d) (g)

Smith, S., D. Simmons, M. Gleason, E. Kameenui, S. Baker, M. Sprick, et al. 2001. An analysis of phonological awareness instruction in four kindergarten basal reading programs. *Reading and Writing Quarterly* 17: 25-50. (d)

Snider, V. E. 1990. Direct Instruction reading with average first-graders. *Reading Improvement* 27: 143-148. (b) (g) (h)

Stein, M. 1990. *Reading research … and Reading Mastery*. Chicago: SRA. (b)

Tobin, K. 2000. *A comparison between Horizons Fast Track A-B and Silver, Burdett, and Ginn reading curricula in first grade: June 2000 final report*. Oregon: National Institute for Direct Instruction. (a) (e)

Torgesen, J., S. Morgan, and C. Davis. 1992. Effects of two types of phonological awareness training on word learning in kindergarten children. *Journal of Educational Psychology* 84: 364-370. (d)

Traweek, D. and Berninger, V. 1997. Comparisons of beginning literacy programs: Alternative paths to the same learning outcome. *Learning Disability Quarterly* 20: 160-168. (a) (d) (h)

Umbach, B. T., C. Darch, and G. Halpin. 1987. *Teaching reading to low performing first graders: A comparison of two instructional approaches*. ERIC 290 130. (b)

Umbach, B., C. Darch, and G. Halpin. 1989. Teaching reading to low performing first graders in rural schools: A comparison of two instructional approaches. *Journal of Instructional Psychology* 16: 23-30. (h)

Urdegar, S. M. 1998. *Evaluation of the Success for All Program, 1997-98*. Miami, FL: Miami-Dade Public Schools, Office of Evaluation Research. (a)

Urdegar,k S. M. 2000. *Evaluation of the Success for All Program, 1998-99*. Miami, FL: Miami-Dade Public Schools, Office of Evaluation Research. (a)

Weinstein, G. and N. L. Cooke. 1992. The effects of two repeated reading interventions on generalization of fluency. *Learning Disability Quarterly* 15: 21-28. (d)

Varela-Russo, C., K. A. Blasik, and M. Ligas. 1998. *Alliance of quality schools evaluation report*. Ft. Lauderdale, FL: School Board of Broward County. (a)

Yu, L. and R. Rachor. 2000. *The two-year evaluation of the three-year Direct Instruction program, in an urban public school system*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 441831). (a) (e)

Note: Letters in parentheses refer to the source(s) providing the citation.
a. Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2003. Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research* 73: 125-230.
b. Adams, Gary L. and Siegfried Engelmann. 1996. *Research on Direct Instruction: 25 years beyond DISTAR*. Seattle: Educational Achievement Systems.
c. McGraw-Hill, 2002. "Appendix: Research supporting instruction in *Reading Mastery*: A Selected Annotated Bibliography." *Results with Reading Mastery*. ERIC ED 469 523.
d. Schieffer, Cheryl, Nancy E. Marchand-Martella, Ronald C. Martella, Flint L. Simonsen, and Kathleen M. Waldron-Soler. 2002. An analysis of the *reading mastery* program: Effective components and research review. *Journal of Direct Instruction* 2: 87-199.
e. Comprehensive School Reform Quality Center. 2006. *CSRQ Center Report on Elementary School Comprehensive School Reform Models*. Washington, D.C. American Institutes for Research.
f. NIFDI research staff found this citation independently.
g. Grossen, Bonnie. n.d. "References for *Reading Mastery* Reading Review," The Research Base for *Reading Mastery*, SRA, Eugene, Oregon: University of Oregon. (http://darkwing.uoregon.edu/~adiep/rmref.htm).
h. National Reading Panel. 2000. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washing ton, D.C.: National Institute of Child Health and Human Development.
i. Kinder, D., R. Kubina, N.E. Marchand-Martella. 2005. Special education and Direct Instruction: An effective combination. *Journal of Direct Instruction* 5: 1-36.

Appendix D
The Impact on Summary Ratings of Ignoring Fidelity of Implementation

As noted in the text of this report (p. 11), the WWC has indicated that they have chosen to ignore ratings of fidelity of implementation and, instead, to rely on "replicated findings, which ensures that any one study in which fidelity issues may have arisen are averaged with findings from other studies" (Dynarski, 2008, p. 3, in Appendix A of this report). As noted in the text, this assumption is seriously flawed. In fact, for both exemplary programs and for programs that are ineffective, poor implementation of a program would, very likely, produce results that are systematically biased. This has very serious consequences for consumers, leading to minimizing the positive impact of good programs as well as the negative impact of poor programs.

To illustrate these results the tables below examine the possible implications of inadequate fidelity of implementation for two different hypothetical curricular programs: one that is more effective than a control group and one that is less effective than the control group. We use effect sizes (the difference between the mean scores of experimental and control groups divided by the common standard deviation) as the metric of comparison. Basic to our analysis is the assumption of "regression toward the mean," the well established statistical phenomenon where those with high scores (or low scores) on a measure will tend to have scores that are closer to the mean (lower for those who are high and higher for those who are low) at later testing periods. Simply through regression toward the mean we would expect those who score lower than a mean at pretest to score higher (and closer to the mean) at posttest. For sake of illustration we will assume that the scores involved are normal curve equivalent (NCE) scores with a control group mean of 50 and a standard deviation of 21.

Consider first a program that is effective. The first column of Table D-1 gives four possible values of the experimental group that we will assume are the "true" values and the second column gives the value of Cohen's d that results when comparing these experimental means with those of the control group. It can be seen that the d values vary from .24, when the experimental mean is 55, to .95, when the experimental mean is 70. The higher values in this column are, in fact, similar to those often obtained in studies of Direct Instruction, especially when it is implemented with fidelity.

The other columns in part A of Table D-1 examine the likely result if the effective program is implemented with less than optimal fidelity, but the impact is simply the "random" effects that the WWC assumes will occur. In these calculations the means of the experimental group do not change, as implied by the WWC assumption. However, the standard deviation is assumed to change, for with lower levels of fidelity, even if the means stay constant, greater variability would be expected. This is at the basis of the WW notion of "averaging out." Three different values of the standard deviation are given, and it can be seen that in all situations the value of the effect size becomes lower than the "true" value, often substantially so. (Compare the "true" values in the second column with the other values of d going across each row.)

Panel B of Table D-1 presents results that are probably more likely if an effective program is implemented with less than adequate fidelity. In such a situation it would be logical to expect that performance would be lower – that the mean value would decline. As with Panel

A, the first two columns of Panel B give the "true" experimental mean and effect size for comparison purposes. The next two columns present the effect sizes that would occur if the average value of the experimental group declined by either 5 points (column 3) or by 10 points), but assuming no change in the standard deviation. Again, all of the effect sizes are smaller. Finally, the calculations in the last three columns of Panel B assume that lower fidelity with an effective program has two effects: the average score is lower than it would otherwise be and the standard deviation is larger, as in Panel A. Again, of course, the effect sizes are substantially lower and the declines are marked.

Table D-1: Effect Sizes of Comparisons with Effective Programs with Different Fidelity Conditions

A. Random Influences of Fidelity Problems

Mean, Experimental Group	"True" d	s.d. = 25	s.d. = 30	s.d. = 35
55	0.24	0.20	0.17	0.14
60	0.47	0.40	0.33	0.29
65	0.71	0.60	0.50	0.43
70	0.95	0.80	0.67	0.57

B. Systematic Influences of Fidelity Problems

		Systematic - Experimental mean smaller, s.d. the same		Systematic - Exp Mean smaller, s.d. larger		
Mean, Experimental Group	"True" d	Mean 5 points less	Mean 10 points less	mean 5 points less, s.d. = 25	mean 10 points less, s.d. = 25	mean down by 10 points, s.d. = 30
55	0.24	0.00	-0.24	0.00	-0.20	-0.17
60	0.47	0.24	0.00	0.20	0.00	0.00
65	0.71	0.47	0.24	0.40	0.20	0.17
70	0.95	0.71	0.47	0.60	0.40	0.33

Note: If the impact of fidelity implementation is random, it is assumed that this affects only the common standard deviation and not the mean. If the impact of fidelity implementation is systematic, both the mean and the standard deviation can be affected. Cohen's d is calculated by subtracting the mean of the experimental group from the mean of the control group and dividing by the common standard deviation. The "true" calculations assume that the standard deviation is 21, the value of the control group = 50 and experimental mean varies as shown in each of the rows of the table.

Now consider the possible implications of poor fidelity of implementation of a program that is, in reality, ineffective and, in fact, produces poorer results than the control group. These results are shown in Table D-2. As with Table D-1, the results with the "true" values are shown

in the first two columns of each panel. It can be seen that the "true" effect sizes vary from -.24 to -.95.

The results in Panel A illustrate what would happen if the fidelity problems produce only random changes: the means would stay the same, but the standard deviations would become larger. It can be seen that, in all cases, the absolute values of the effect sizes become smaller – that is, less negative.

Table D-2: Effect Sizes of Comparisons with Ineffective Programs with Different Fidelity Conditions

A. Random Influences of Fidelity Problems

Mean, Experimental Group	"True" d	s.d. = 25	s.d. = 30	s.d. = 35
45	-0.24	-0.20	-0.17	-0.14
40	-0.47	-0.40	-0.33	-0.29
35	-0.71	-0.60	-0.50	-0.43
30	-0.95	-0.80	-0.67	-0.57

B. Systematic Influences of Fidelity Problems

		Systematic - Experimental mean larger, s.d. the same		Systematic - Experimental Mean higher, s.d. smaller		
Mean, Experimental Group	"True" d	mean 5 points higher	mean 10 points higher	mean 5 points higher, s.d. = 18	mean 10 points higher, s.d. = 18	mean 10 points higher, s.d. = 15
45	-0.24	0.00	0.24	0.00	0.28	0.33
40	-0.47	-0.24	0.00	-0.28	0.00	0.00
35	-0.71	-0.47	-0.24	-0.56	-0.28	-0.33
30	-0.95	-0.71	-0.47	-0.83	-0.56	-0.67

Note: If the impact of fidelity implementation is random, it is assumed that this affects only the common standard deviation and not the mean. If the impact of fidelity implementation is systematic, both the mean and the standard deviation can be affected. Cohen's d is calculated by subtracting the mean of the experimental group from the mean of the control group and dividing by the common standard deviation. The "true" calculations assume that the standard deviation is 21, the value of the experimental mean varies as shown in each of the rows of the table.

The results in Panel B illustrate the results if the impact of poor fidelity is systematic. With programs that are ineffective, it would be expected that poor implementation would lead to higher average scores (through regression to the mean) and smaller standard deviations. The

smaller standard deviations result from having values that are "less ineffective." As can be seen in Table D-2, the result is that the effect sizes are less negative. That is, the extent to which the programs are truly ineffective, as shown by the "true" values of d, is disguised.

The results presented above seem to clearly refute the assumption that guides the WWC's approach to considering the fidelity of implementation in deliberations. Even if the results of poor fidelity of implementation were random, the efficacy of both good and poor programs would be misrepresented. If, as is more likely, the results of poor implementation are not random, the impact would be even greater. Good programs appear less effective and poor programs appear better than they actually are. The WWC's decision regarding consideration of the fidelity of implementation appears to be seriously misguided.