# *Reading Mastery* for Beginning Readers: An Analysis of Errors in a What Works Clearinghouse Report

Technical Report 2014-4

Jean Stockard, Director of Research and Evaluation, NIFDI

August 25, 2014

# Table of Contents

# List of Tables

# Executive Summary[1]

The scholarly literature includes dozens of studies that show that the Direct Instruction program, *Reading Mastery (RM)*, is highly effective. Despite this large and consistent body of work, a report published by the What Works Clearinghouse in November 2013 stated that it could find "no studies of *Reading Mastery* that fall within the scope of the Beginning Reading review protocol [and] meet What Works Clearinghouse (WWC) evidence standards" (WWC, 2013b, p. 1). This NIFDI technical report documents a surprising number of errors in the WWC analysis.

The WWC uses a three step process to identify research studies of educational interventions. Significant errors were found at each stage. The first step identifies studies for review. While the WWC 2013 report listed 166 studies found at this step, *fifteen of them (9% of the total) were either not about Reading Mastery or were included twice in the list. In addition, forty-two efficacy studies of RM that clearly met the protocol were omitted from the list.* Fifteen of these articles were included in bibliographies of relevant articles submitted to the WWC by NIFDI,

At step two the WWC screens studies in two general areas. The first area of review is designed to ensure that the studies meet the review protocol in terms of subject matter, range of grades, student population, and date of publication identified for the review. The WWC reported that 60 studies (omitting the non-*RM* studies) were rejected for these reasons. NIFDI's analysis indicated that 76 should have been rejected for these reasons. Thus, of the 150 unique studies of Reading Mastery identified by the WWC only 74 were efficacy studies that met the review protocol. The 74 efficacy studies identified by the WWC and the 42 additional studies identified by the author could be seen as a relatively complete listing of the literature on *Reading Mastery* for beginning readers in general education that met the review protocol. *The WWC did not identify over one-third of this literature.*

The second area of review at step two involves screening the efficacy studies to see if they use a pretest-posttest control group design, which the WWC requires. Of the 74 efficacy studies that were identified by the WWC, 43 were rejected from consideration at step 2 for having an unacceptable design. As explained in the companion technical report, standard methodological writings in the social sciences note that a wide variety of other designs are just as good, and often better, for educational settings, than the one preferred by the WWC. *If standard methodological practices regarding acceptable research designs were used,*

[1] The author gratefully acknowledges the extraordinarily skilled assistance of Timothy Wood in compiling information for this report and the helpful comments of Muriel Berkeley, Douglas Carnine, Christina Cox, Gary Davis, Siegfried Engelmann, and Jerry Silbert on earlier drafts. All conclusions and opinions in this document are, however, the sole responsibility of the author.

*about three-fifths of the studies rejected at step 2 for reasons related to design could have been examined.*

The third step of the WWC review process is designed to see if the studies accepted at step two conform to WWC "standards of evidence." At this stage, the WWC rejected all of the remaining 31 efficacy studies in their list. The companion technical report compares the WWC standards of evidence to accepted methodological practices, documenting significant differences in the WWC approach and the scholarly literature. *Analysis of the 31 studies rejected at step 3 determined that, if standard methodological criteria had been used, almost half of them (n=15) would have been included in a review.*

In total, of the 74 efficacy studies that the WWC identified as fitting the review protocol for Beginning Reading, over half (n=40 or 54%) would have been examined if the traditionally accepted methodological standards of the social sciences had been applied. Each of these studies employed a comparison group design that is commonly accepted within the methodological literature. *When the 42 studies not included in the WWC listing are considered, there are over 80 studies of RM that fit the WWC's protocol, used a comparison group design, met standard methodological criteria for internally valid research designs, and thus could have been reviewed.*

Effect sizes (Cohen's d) were computed for results of 38 of the 40 analyses identified by the WWC that could have been examined. Effect sizes summarize the difference between an intervention and a control condition as a percentage of the common standard deviation (variability) of the two groups. An effect size of zero indicates no difference between the groups. Traditionally, effect sizes of .25 or larger have been seen as educationally significant (Tallmadge, 1977). The average effect sizes in these 38 analyses ranged from -.53 to 2.44. Almost three fourths of the studies had average effects that were larger than the .25 criterion. Only one had an average effect less than -.25 (showing an educationally significant advantage for the non-RM program). *The average effect across all 38 studies was .57, more than twice the .25 level denoting educational importance. This average value is similar to the average effect sizes found in meta-analyses of RM in the research literature.*

It is suggested that substantial changes are needed in WWC procedures and policies to ensure that reports accurately reflect the research literature. Until such changes are made users are advised to consult reviews of studies in the standard research literature rather than the WWC summaries.

# Reading Mastery for Beginning Readers: An Analysis of Errors in a What Works Clearinghouse Report

The What Works Clearinghouse (WWC) is a federally funded program established in 2002 to evaluate educational interventions and provide summary ratings and reports of their effectiveness. The WWC's website describes their organization as a "trusted source of scientific evidence for what works in education to improve student outcomes" and as providing "accurate information on education research" (WWC, 2013a). Yet, the reports of the WWC often directly contradict those within the research literature, giving high ratings to programs for which scholars have found little positive evidence and negative ratings to programs that have accrued substantial positive evidence. Such contradictions are associated with a report on *Reading Mastery* issued by the WWC in November, 2013.

*Reading Mastery (RM)* is a reading program that is part of the Direct Instruction corpus of curricula. A large body of literature has documented the effectiveness of Direct Instruction programs in promoting achievement. In a recent review Coughlin (2014) summarized several extensive reviews of these studies, noting the consistent and strong support for the programs' efficacy (e.g. Adams and Engelmann, 1996; Kinder, et al., 2005; Przychodzin-Havis, et al., 2005; Schieffer, et al., 2002). Authors of the meta-analyses have commented on the extensive nature and quality of this literature base, especially in comparison to other programs (e.g Borman, et al., 2003, p. 141; Hattie, 2009, pp. 206-207). Yet, the WWC's November 2013 review reported that it could find "no studies of *Reading Mastery* that fall within the scope of the Beginning Reading review protocol [and] meet What Works Clearinghouse (WWC) evidence standards" (WWC, 2013b, p. 1). Why did this discrepancy occur? Given the large literature about *Reading Mastery*, how could the WWC find no studies of its efficacy that it considered acceptable?

This paper, and a companion one (NIFDI Technical Report 2014-3) examine that question. The results indicate that the differences involve two general issues: 1) the policies of the WWC and the ways in which they differ from those typically used within the social sciences and 2) errors in the review procedures of the WWC. The companion report examines the policies of the WWC, contrasting their procedures with traditional and standard methodological approaches. This report focuses specifically on errors in the November 2013 analysis of *Reading Mastery*. A forthcoming report (NIFDI Technical Report 2014-5) documents the extent to which numerous other organizations and individuals have found serious errors in WWC procedures, policies, and reports.

The WWC's *Procedures and Standards Handbook* (WWC, 2014) describes three steps of work in selecting studies to include in reviews. The first involves identifying relevant literature. This is followed by an initial screening of the identified materials to determine if they are eligible for review within the protocol and if they use the required type of research design. Studies that meet these general eligibility criteria are then examined to see if they meet the WWC methodological standards of evidence. Only studies that pass this third step are then reviewed. For the November 2013 report none of the identified studies passed this third step. However, our analysis found disturbing errors in each of the points in the decision process. If more appropriate review procedures had been used, the conclusions regarding *Reading Mastery* would have been strikingly different and would have paralleled the judgment of the scholarly research community that *Reading Mastery* is highly effective.

The following sections of this report examine results of the three steps in the WWC's 2013 review of *Reading Mastery* for beginning readers, documenting errors that occurred at each step. A final section summarizes the analysis and discusses ways in which errors such as those documented here could be avoided in the future. Extensive appendices provide supporting details.

## Step 1: Identifying Relevant Literature

Summaries of the literature always begin with identifying relevant sources. The WWC protocol for the review of beginning reading programs (WWC 2012) describes its procedures for this step, listing databases to be consulted, keywords, and other sources. The general WWC procedures and policy handbook notes that

> studies are gathered through a comprehensive search of published and unpublished research literature, including submissions from intervention distributors/developers, researchers, and the public to the WWC Help Desk. Only studies that are publicly available are eligible for inclusion in a WWC review (WWC, 2014, p. 6)

From their review of the Direct Instruction literature the WWC reported that it "identified 166 studies of *Reading Mastery* for beginning readers that were published or released between 1983 and 2012" (WWC 2013b, p. 1) (Citations to all studies included in the WWC listing are in Appendix A.)

Examination of this listing found surprising errors. One of the most disturbing results of the review was the discovery that the list of 166 studies actually included 13 works that did not involve *Reading Mastery*, but instead focused on other programs. Five of the studies looked at Reading Recovery, an individualized tutoring program that uses a whole language approach and is not related at all to the Direct Instruction tradition to which *Reading*

*Mastery* belongs.[2] Six other studies involved a variety of approaches, such as "Big Books," Headsprout, and Earobics;[3] and two used other Direct Instruction programs, *Corrective Reading* and *Horizons*.[4] The fact that these studies did not examine *Reading Mastery* is usually clear from the titles of the works, and is always apparent from the content.[5] Thus, the WWC actually identified 153 studies involving *RM*, not 166. Approximately eight percent of the studies originally listed did not involve Reading Mastery, but were about other programs.

Two works were listed twice by the WWC. One was written by SRA/McGraw Hill (2005, n.d.m) and appeared with the same title and author but different years. Another study (O'Brien and Ware, 2002) was given two different designations. It was listed as ineligible for review because it did not use an accepted study design (rejected at step 2), However, it was also included in the list of studies that passed the protocol screening at step 2, but was rejected at step 3 for using "a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent" (WWC 2013b, pp. 4, 10).[6] Two other articles (Marchand-Martella, et al., 2006, 2007) reported on the same data set with identical results given in both publications. These two articles were only considered once in the present analysis. When the two SRA/McGraw Hill articles and the Marchand-Martella works are combined, 150 unique studies related to *RM* were in the list of studies originally identified by the WWC. (See Table 1.)

The NIFDI Office of Research and Evaluation notified the WWC of a number of these errors in early 2014. The WWC issued a revised report in March, 2014 that corrected the errors that NIFDI had reported to them, noting that they had been mistakenly included. However, the March 2014 posting retained the date of the WWC report as November, 2013, giving readers no discernable indication that there was new information or revisions included in the report.

---

[2] European Centre for Reading Recovery, 2012; Harvey, 2012; Lewis, 2012; McClendon, 2012; Redding, 2012.

[3] Cohen & Brady, 2011; Hudler, 2008; Keafer, 2008; Laska, 2009; Schelling, 2010; Welsh, 2010.

[4] Flores & Ganz, 2007; SRA/McGraw Hill, 2005n.d.l.

[5] Three of the thirteen studies were found to be eligible for review (meeting the protocol for the review and examined, and subsequently rejected, at stage 3). Of these, two (McClendon, 2012; Redding, 2012) were rejected because the groups were found to be unequal prior to the intervention, and one (Harvey, 2012) was rejected because the outcomes were measured in a way that was inconsistent with the protocol. The other studies were reported as not meeting the eligibility criteria and were rejected at stage 2. Lewis (2012) was omitted because fewer than 50% of the students were in general education. The other 9 articles were rejected because they did not use one of the accepted designs.

[6] While the WWC includes no explanation of this discrepancy it may reflect the fact that two different analysis procedures were used in the O'Brien and Ware (2002) study, as described in Appendix B. Another study (Stockard and Engelmann, 2010) was also listed twice, but this reflected the fact that it reported data on two different samples each involving different designs. The WWC made this distinction in their listing. One part of the Stockard and Engelmann (2010) study used a subset of the data included in Stockard (2011). These results are combined in the computations of average effect sizes and the meta-analysis in the companion technical report. For purposes of counting articles they have remained separate.

Table 1

*Errors in Decisions at Step 1 of the Selection of Studies for the November 2013 WWC Review of Reading Mastery*

| | |
|---|---|
| Number of Studies in Original List | 166 |
| Number of Studies About Programs Other than RM | 13 |
| Number of Studies Listed Twice | 3 |
| Corrected Number of Unique Studies of RM Identified | 150 |

The WWC review also omitted a relatively large number of studies about the program. The review protocol states that the literature search will include "checking prior reviews and research syntheses (i.e. using the reference lists of prior reviews and research syntheses to make sure that key studies have not been omitted)" (WWC, 2012, p. 19) . To date, 42 studies of *RM* that appear to fit the protocol and should have been included in the listing have been identified. Over a third (n=15) of these works had been given to the WWC in NIFDI's response to their 2008 report on *Reading Mastery*. It is unclear why the WWC chose to ignore the bibliography that had been supplied to them, for, as cited above, the procedures handbook notes explicitly that such submissions will be considered.[7] The list of articles that were omitted and the procedure for choosing them is in Appendix B. It is important to note that these 42 studies are a relatively small subset of the available studies on *Reading Mastery*. They include only those that appear to fit the protocol and would have passed the screening process in step 2, which is described in the next section.

## Step 2: Screening Materials for Eligibility

The second step of the WWC review process involves determining if the identified studies met the protocol for the review and were thus eligible for further screening. At this step the WWC rejected about 80 percent of the studies (132 of the 166 studies listed, or 120 of the 150 in the corrected count). Again, there were surprising discrepancies and apparent errors in these decisions. The errors "went both ways." In other words, sometimes the WWC included studies for review when they clearly did not meet the protocol. At other times, they rejected studies when they could have, arguably, been included.

[7] An analysis of the WWC review of the use of *Reading Mastery* with students with learning disabilities documented a similar pattern of the WWC ignoring a large proportion of the literature, even when it had been supplied to them. As described by Stockard and Wood (2013, p. 4), "the WWC appears to have ignored well over half of the material that directly addressed the use of *Reading Mastery* with students with learning disabilities, even when presented with extensive analyses that demonstrated its relevance. Moreover, their decisions about which studies to include appear to be highly selective."

Table Two summarizes the reasons that the WWC used to discard studies at this step. The reasons for rejecting a study are separated into two broad categories: 1) those not related to the design of a study (top part of the table) and 2) those related to a study's design (bottom part of the table). The first column of data gives the numbers reported by the WWC (n = 132). The second column gives the numbers when studies regarding programs other than *RM* and the double counted items were counted only once (n=120).[8] The third column summarizes the analysis of the material discussed below regarding the number of studies for which the WWC's decisions could be questioned and the determination of a more accurate classification of the 120 studies that were rejected at step 2.

From their review of the identified studies the WWC staff concluded that 60 studies were ineligible for reasons not related directly to their design. Two general types of reasons were given: 1) the studies did not analyze the effect of *RM*, but instead were a literature review, meta-analysis, or other type of study (n=22, corrected count in column 2); or 2) they did not conform to the protocol's criteria regarding the specified grade range (K-3 for beginning reading), student characteristics (general education), or outcome measures (reading skills) (n=38, corrected count in column 2). Because of time constraints, a complete review of the 60 articles that were rejected for these non-design related reasons was not conducted. (None of the 42 additional studies that are in Appendix B would have been rejected for these reasons.) However, each of the 60 articles rejected for reasons related to the study's design was examined. A summary of the errors found in these decisions is discussed in the first sub-section below. The second sub-section describes the results of studies that should, arguably, have been included in the WWC's analysis if standard methodological criteria had been used.

## Decision Errors at Step Two

Of the 60 studies rejected at step 2 for reasons related to the study's design, most were discarded because they did not use a "comparison group design" (n=58), and the others were omitted because the intervention was judged to be bundled with other components (n=2). Examination of these 60 studies revealed a number of cases where these decisions could be questioned.

First, the examination indicated that 16 of these 60 studies did not fit the protocol and should have been excluded for non-design related reasons. Two of the reports were not

[8]As explained in the text, Marchand-Martella, et al. 2006 and Marchand-Martella, et al., 2007 reported on the same data set and were treated as only one study in our analysis. The O'Brien and Ware (2002) article has remained as part of the count in our analyses of both Stage 2 and Stage 3 because it reported on two types of designs, one of which may have been more acceptable to the WWC than the other. (It was reported as rejected in both stages.)

efficacy studies, but looked at implementation and teacher attitudes;[9] eight did not look at general education students, but at programs serving students with moderate to severe disabilities;[10] and four did not disaggregate and/or include data for grades included in the WWC's Beginning Reading Protocol.[11] In addition, despite extensive searches we were never able to locate copies of two works and are unsure if they actually exist.[12]

Table Two

*Studies Rejected at Step 2 of the WWC Review (n=120, Corrected Count)*

| | Original WWC Count | When Non-*RM* Studies and Duplicates are Omitted | Corrected Numbers Based on Our Review |
|---|---|---|---|
| Reasons Not Related to Study Design | | | |
| Literature review or meta-analysis | 15 | 15 | 15 |
| Did not examine effectiveness of an intervention | 7 | 7 | 9 |
| Students were not in Grades K-3 | 19 | 19 | 19 |
| Fewer than 50% General Education Students | 9 | 8 | 16 |
| Findings not disaggregated by Age or Grade Range Specified | 8 | 8 | 12 |
| Did not include a domain specified in the protocol | 3 | 3 | 3 |
| Not publicly available | 0 | 0 | 2 |
| Subtotal | 61 | 60 | 76 |
| Reasons Related to Study Design | | | |
| Bundled the intervention with other components | 2 | 2 | 10 |
| Does not use a "comparison group design" | 69 | 58 | 9 |
| Design Acceptable with Standard Criteria | 0 | 0 | 25 |
| Subtotal | 71 | 60 | 44 |
| Total | 132 | 120 | 120 |

Note: Discrepancies between the first and second column of data reflect the original inclusion of 10 articles that were not about *RM*, two that were double counted (SRA/McGraw Hill 2005j, n.d.m.), and two articles that reported on the same data set (Marchand-Martella, et al., 2006, 2007). When the 10 non-RM articles and two of the double-counted ones are removed the total number rejected at step 2 is 120 rather than 132. (Another of the double counted articles (O'Brien and Ware) was retained in this section because it also appeared in decisions regarding step 3. In addition, three of the non-RM studies were not rejected until step 3 of the review.)

[9] Elias, 2009; Shelton, 2010

[10] Chamberlain, 1987; De La Cruz, 2009; Humphries, et al., 2005; SRA/McGraw-Hill 2006f,g; 2007c, n.d. h, o

[11] SRA/McGraw Hill 2006c,d; 2007a;n.d.f

[12] Asfendis, 2008; Intensive, tailored tuition raise literacy, 2012. Searches in WorldCat, an index of material in libraries throughout the world, did not include either of these works. A search for the 2012 article in the journal in which it was supposedly published held nothing related to the article title nor any entries for Reading Mastery nor Direct Instruction.

Second, using standard methodological criteria there are reasons to question a number of the WWC's decisions to reject studies. While the WWC reported that two studies "bundled" *RM* with another intervention, this issue actually affected ten of the listed studies (one of those listed by the WWC in this category and nine additional studies[13]). The WWC reported that all the other efficacy studies were rejected because they "did not use a comparison group design." This decision appears to be accurate for 9 of the studies,[14] but could be legitimately questioned for the others. In total, there appear to be reasons to question over half of the WWC's decisions regarding studies rejected at step two for reasons related to the study design, a surprisingly high rate of error. The errors appeared to occur with each of the types of decisions that were made. The next section summarizes the results of studies that should have, arguably, been considered if standard methodological criteria had been used.

## Studies of RM That Could Have Been Accepted At Step Two but Were Not

Appendix C describes each of the 25 studies that were rejected at step 2 but that would have been accepted for review by the WWC if standard methodological criteria had been used. Methodologists recommend a large number of research designs for studies in field settings such as schools. As described in the companion report (NIFDI Technical Report 2014-3), the standard literature specifically cautions against the use of randomized control trials in such settings and recommends a variety of other designs that can have high internal and external validity. It also calls for the accumulation of many different studies and comparing the results across a range of populations and settings. The discussion in this sub-section builds on that tradition.

Each of the 25 efficacy studies identified by the WWC and rejected at step 2 that should have arguably been included for review had data on a comparative population that did not receive *RM*. This allows the computation of effect sizes, a standard metric used to describe the impact of an intervention. Most of the studies used a variation of the cohort control group design, a design specifically recommended as appropriate for school settings (Campbell and Stanley, 1963, Cook and Campbell, 1979; Shadish, Cook, and Campbell, 2002). Of the 20 studies that used this design, eight used data from previous cohorts as a comparison group, nine compared changes across cohorts to those in a larger population, and three compared changes across cohorts to normative populations. Other designs used were a pretest-posttest normative control group design (n=2), posttest only design with the control and intervention groups matched on pre-intervention achievement (n=1), a pretest-posttest gain score repeated measures design (n=1), and a pretest-posttest control group

[13] Kubinda, Commons, & Heckard, 2009; Simmons, Coyne, Kwok, McDonagh, Harn, & Kame'enui, 2008; Nanda & Fredrick, 2007; SRA/McGraw Hill, 2007e; n.d. c, e, i, j, k, q

[14] SRA/McGraw Hill 2005a,c,f,g,k, m; 2006h; n.d.d, g

design with random assignment (n=1).[15] Most importantly, all of these studies have designs that incorporate "comparison groups" and, as documented in the companion report, are recommended by standard methodological traditions in the social sciences.

Effect sizes (Cohen's *d*) were calculated for 24 of the 25 studies.[16] Effect sizes are a common metric used to summarize the impact of educational interventions. They summarize the difference between an intervention and a control condition as a percentage of the common standard deviation (variability) of the two groups. An effect size of zero indicates no difference between the groups. Traditionally, effect sizes of .25 or larger have been seen as educationally important (Tallmadge, 1977). Only one of the 24 studies had an average effect size that was negative. In contrast, 19 of the studies (79%) had average effect sizes that equaled or surpassed .25. The average effect size was .54, more than twice the traditional standard.[17] This value is similar to that reported in various meta-analyses of the impact of *Reading Mastery* on student achievement. Descriptions of each of these 25 studies are in Appendix C.

## Step 3: Meeting the WWC Standards of Evidence

The WWC reported that 31 of the 150 studies of RM that were identified (corrected count) employed an acceptable design and were thus examined at step 3 to see if they met the WWC standards of evidence. In other words, the WWC found 31 studies that employed a pretest-posttest control group design (passing step 2). However, they found that none of these studies met the additional "standards of evidence." Thus, they were all excluded from consideration at step 3.[18]

This section looks at these 31 studies and the reasons that they were rejected for further analysis. As shown in Table Three, four general reasons were cited: a lack of group equivalence at pretest (n=13), unacceptable confound (n=10), having "one unit per condition" (n=7), and not meeting WWC pilot single-case design standards (n=1). The companion report has a methodologically oriented discussion of these standards. This report focuses on the validity of the decisions.

---

[15] This large federally funded study would appear to meet the WWC's definition of an acceptable design and has been accepted in WWC reviews of other topic areas. It was rejected for the November 2013 review of *RM* for beginning reading because of a supposed bundling of *RM* with other interventions. Details on the study and a specific critique of the WWC's decision are in the discussion of this study in Appendix C.

[16] At the time of writing we had not been able to locate information on norms for one of the studies (Watkins, 2008), so it is not included in the analysis of effect sizes.

[17] This was calculated using the study as the unit of analysis. The companion technical report gives results of mixed models that had the individual effect size as the unit of analysis and controlled for study level effects.

[18] As described above, three other studies were listed in the original WWC report about *RM* as being rejected at step 3, but actually involved a curriculum other than *Reading Mastery* (Harvey, 2012; McClendon, 2012; Redding, 2012).The studies of McClendon (2012) and Redding (2012) were rejected because the groups were deemed to be not equivalent at baseline. Harvey (2012) was rejected for having "outcomes that are overaligned with the intervention or measured in a way that is inconsistent with the protocol." It is unclear why reviewers were able to make these judgments without understanding that the studies did not involve *RM*.

Table Three
*Studies Rejected at Step 3 of the WWC Review (n=31, Corrected Count)*

| Decision | Original WWC Count | When Non-*RM* Studies and Duplicates are Omitted | Corrected Numbers Based on Our Review |
|---|---|---|---|
| Non-Equivalent Groups | 15 | 13 | 7 |
| Confounded Design | 10 | 10 | 6 |
| One Unit Per Condition | 7 | 7 | 3 |
| Doesn't Meet Case Study Criteria | 1 | 1 | 0 |
| Problematic Measure | 1 | 0 | 0 |
| Eligible for Review | 0 | 0 | 15 |
| Total | 34 | 31 | 31 |

Note: The discrepancies between the first and second columns of data reflect the inclusion of three studies that were about programs other than *Reading Mastery* in the WWC analysis. Two of these were rejected for having non-equivalent groups and the third was rejected for issues with the dependent measure. It is unclear why the reviewers did not realize that the studies involved a program other than *Reading Mastery*.

As with the decisions at step 2, examination indicated that the WWC's decisions appeared appropriate in some cases, but, if criteria commonly accepted in the social sciences were used, almost half of these studies (n=15) should have been included for analysis. Appendix D has descriptions of each of the 31 studies and the effect sizes associated with the results of those that should have, arguably, been retained for analysis. The discussion in the following sub-sections summarizes the analysis, looking at each of the general WWC standards.

## Studies Rejected for a Lack of Group Equivalence

Thirteen of the 31 studies rejected at step 3 were dismissed from consideration by the WWC because they used "a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent." In other words, while the design compared two groups, the WWC determined that the groups were not equivalent before the intervention occurred. Our analysis indicated that seven of these 13 decisions were appropriate. These studies included various characteristics that rendered the groups incomparable including combining *RM* with another curriculum ;[19] including only students who received *RM*, but different time devoted to instruction;[20] and differing levels of

[19] Kamps, et al, 2008, Neely, 1995; Ryder, et al., 2003; Thames, et al., 2006;Thomson, 1991
[20] League, 2001

implementation support and/or no pretest information[21] (see Appendix D for additional details). However, this decision could be questioned with the other six studies within this sub-group.[22] All of these studies used a pretest-posttest control group design, and all but one of them used multivariate statistics to control for pre-test differences. The exception used gain scores as a method of control.

The WWC has stringent, set criteria regarding differences between groups at pretest, rejecting any study for which the differences exceed .25 of a standard deviation on any of the measures and requiring statistical adjustments if differences are greater than .05 of a standard deviation.[23] Data were provided to examine differences in group scores at pretest for four of the six studies. Our calculations indicated that all of the differences fell well below the .25 criterion set by the WWC (and thus should have been acceptable with statistical controls) for three of these four articles. It is unclear why these three studies were rejected using this criterion. The exception involved a very large federally funded study (Crowe, et al., 2009) that included comparisons with several different curricula and across multiple grades. The intervention group differed at pretest from the comparison group by more than .25 s.d. on 3 of the 15 comparisons, but the average difference between the groups was .12 s.d. As explained in the companion report, given the number of comparisons involved, the few differences that exceeded the .25 criterion would be quite likely to have appeared simply by chance.

The effect sizes (Cohen's d) associated with these six studies ranged from .11 to .97, with an average of .42. Three-fifths of the effects were larger than the .25 level generally used to denote educational importance.

## Studies Rejected for an Unacceptable Confound

Ten articles were listed as rejected because there was some type of confound within the design: the intervention was combined with another intervention (n=8), the effects were not reported separately for the intervention (n=1), or the "intervention was not implemented as

[21] McCollum, et al., 2007

[22] Two of the six studies (Stockard, 2011 and Stockard and Engelmann, 2010) reported on the same data set, and their results were combined in the analysis reported here and in the multivariate analysis in the companion technical report.

[23] As explained more fully in the companion report, this criterion is relatively difficult to meet, especially when samples are small or when multiple measures are used. Many of the designs recommended by the standard methodological tradition do not require pretests, using other methods to determine group equivalence and internal validity. In addition, the WWC includes no provision for examining studies where an intervention begins a study with substantially lower scores than a comparison group, but, at the end of the intervention have higher scores – clearly indicating an important effect. Several of the studies rejected by the WWC and described in Appendices C and D had this result.

designed" (n=1), an apparent reference to the fidelity of implementation. As described in Appendix D, our analysis indicated that six of these decisions were probably appropriate.[24]

However, the WWC's decision to exclude studies because of potentially confounding effects could be questioned for the other four studies in this group. Two of these were relatively small studies using a pretest-posttest control group design with random assignment. The other two reported evaluations of large-scale implementations in urban areas using a pretest-posttest control group design with statistical adjustments. As described in Appendix D, it is difficult to understand the reason for the WWC's determination regarding the two studies with randomized assignment (Jones, 2002; Umbac, et al., 1989), and it is asserted that the WWC's judgment of one of the large scale field studies (Carlson & Francis, 2002) involved faulty logic. The fourth study (Mac Iver & Kemper, 2002) was rejected because of apparent issues with fidelity of implementation, yet the authors reported data separately for the groups with varying levels of fidelity and those results could have been potentially included in a more inclusive review.[25]

The average effect sizes (Cohen's d) associated with these four studies ranged from .11 to 2.43, with an average of .96. Three fifths of the individual calculated effects were larger than the .25 level generally used to denote educational importance.

### Studies Rejected for Having One Unit per Condition

Seven articles were listed as being rejected because "the measures of effectiveness cannot be attributed solely to the intervention—there was only one unit assigned to one or both conditions." This standard requires that a study include more than one unit in each comparison group, defining a unit as "a single teacher, classroom, school, or district" (WWC, 2014, p. 19) Examination of the seven studies rejected for this reason indicated that three of the decisions appear to have been appropriate.[26] In two of the studies (Ryder, et al, 2006; Wiltz & Wilson, 2006), all of the students received *RM* for reading instruction and there was no comparison group that had not received the program. In the other study, the results with *RM* were combined with other interventions, and the impact of *RM* could not be separately determined (Wills, et al., 2010). Thus, while it was appropriate to exclude the three studies, it does not appear that the "one unit" standard captures the reasons for the exclusion.

The decisions to reject the other four studies could be questioned. One (Ashworth, 1999) compared reading achievement of two classes of second grade students, both of which were taught by the same teacher, but in different years, one class using a basal textbook and the other using RM. Two studies (Green, 2010; Stockard and Engelmann, 2010) used a pretest-

[24] Algozzine, et al., 2012; Foorman, et al., 2003; Kamps, et al., 2007; Kamps & Greenwood, 2005; O'Connor, et al., 2005; and Trout, et al., 2003.
[25] As would be expected, effect sizes were larger with higher levels of fidelity.
[26] Ryder, et al, 2006; Wiltz & Wilson, 2006; Wills, et al., 2010

posttest control group design with statistical controls to compare changes in achievement over time in two schools with similar demographic and achievement histories. The fourth study involved one school, but used a cohort control group design as well as a pretest-posttest with norm comparison design. Each of the latter three studies involved multiple classrooms and teachers in both the intervention and control groups. While all four of the studies failed to meet the WWC's "one unit" standard, by having one teacher or school per treatment, they would be fully acceptable using standard methodological criteria. Studies with one school per treatment had multiple teachers, and the study involving one teacher had different classrooms, a commonly accepted way to control for teacher effects. (See Appendix D for more details).

One of the four studies (Green, 2010) had an average effect size that was negative (-.53), while all the other averages were positive, ranging from .23 to 1.60. The overall average for these four studies was .62, again more than twice the traditional standard for educational importance.

## Study Rejected for Not Meeting WWC Pilot Single-Case Design Standards

One study (Goss & Brown-Chidsey, 2012) was listed as not meeting WWC single-case design standards This study included 6 dyads of first grade students, all in Tier 2 and thus in need of extra help. Their progress in reading through the fall of one school year was examined. Members of the dyads were matched on initial fall DIBELS scores and then randomly assigned to receive Tier 2 instruction in *Reading Mastery* or an alternative program. Several DIBELS measures were collected throughout the fall, and differences in the growth of members of each pair were examined. While all students made progress, those in RM made more progress than those in the other program. The authors summarized their results as follows:

> Of the *RM* students, 4 reached the year-end goal of 50 on NWF in November. By comparison, none of the FDD students [those in the alternate program] met this goal, and only 2 of the FDD students met the adjusted November benchmark goal. These results suggest that *RM* demonstrated significant gains in a short period of intervention time. (p. 70).

The WWC rejected this study "because it does not have at least three attempts to demonstrate an intervention effect at three different points in time." The researchers had one base-line measurement, eight measures within the treatment phase and two follow-up measures. One could well argue that having only one baseline measure was appropriate, given the schools' desire to begin treatment as soon as possible and the matching of students in the two groups. Gathering more baseline measures would have meant that students were denied the extra help they needed and clearly questionable on ethical grounds. A common assumption in the single-subject literature is that multiple baseline

measures are not needed when a researcher can assume that the baselines will remain stable over time. That is logical in this case where reading skills would logically only improve with instruction (Horner & Baer, 1978). In addition, even though there were only 2 later measures, the presence of multiple treatment phase measures would typically be seen as appropriate evidence to consider. If the conclusions of the authors were not supported by the treatment phase measures, exclusion might be warranted. But that was clearly not the case with this study.[27]

## Summary

The scholarly literature includes dozens of studies that show that the Direct Instruction program, *Reading Mastery (RM)*, is highly effective for beginning readers in the primary grades. Despite this large and consistent body of work, a report published by the What Works Clearinghouse in November 2013 stated that it could find "no studies of *Reading Mastery* that fall within the scope of the Beginning Reading review protocol [and] meet What Works Clearinghouse (WWC) evidence standards" (WWC, 2013b, p. 1). This technical report and a companion report (NIFDI Technical Report 2014-3) examine why the WWC's conclusion differs so markedly from the extant literature. The companion report analyzes the WWC's policies and standards. It concludes that the WWC procedures differ markedly from standard procedures used in the social sciences and that the WWC procedures do not provide more accurate estimates of a program's effectiveness. This report examines the WWC decisions used to develop the November 2013 report.

The WWC uses a three step process to identify research studies of educational interventions. Significant errors were found at each stage of the WWC process. This document documents the errors. Standard methodological procedures are used to examine the literature identified by the WWC. The results parallel those reported in other examinations of the literature on *RM*, concluding that the program is highly effective.

## Errors at Step One

The first step identifies studies for review and is described as "systematic and comprehensive" in nature (WWC, 2014, p.4). Yet, our analysis found that numerous studies were included in the listing that should not have been included and other pertinent studies were omitted. While the original WWC 2013 report listed 166 studies found at this step, sixteen of them (9.6% of the total) were either not about *Reading Mastery* or were included twice in the list. Forty-two efficacy studies of *RM* that clearly met the protocol, fifteen of which had been earlier submitted to the WWC in correspondence with NIFDI, were omitted from the list. (See Appendix B.)

[27] Note, in addition, however, that all students in the sample, while in general education, were receiving extra help. In other words, one could argue that the students were not "general education." It is not clear if the study would have been rejected by the WWC for that reason as well. If the WWC were to reject the study for this reason, the rejection should have occurred at Stage 2, not Stage 3.

## Errors at Step Two

At step two the WWC screens studies in two general areas. The first area of review is designed to ensure that the studies meet the review protocol in terms of subject matter, range of grades, student population, and date of publication identified for the review. For example, the November 2013 review was limited to studies of reading skills for general education students in the primary grades published in 1983 or later. The WWC reported that 60 studies (omitting the non-*RM* and double listed studies) were rejected because they did not fit this protocol. Our analysis, however, indicated that somewhat more (76) should have been rejected for these reasons. Thus, of the 150 unique studies of *Reading Mastery* identified by the WWC only 74 were efficacy studies that met the review protocol.

The 74 efficacy studies identified by the WWC and the 42 additional studies listed in Appendix B could be seen as a relatively complete listing of the literature on *Reading Mastery* for beginning readers in general education published after 1983 – a total of 116 studies. The fact that the WWC failed to find over one-third (42/116=36.2%) of this literature could be seen as disturbing and casts considerable doubt on the assertion that their search of the literature was "comprehensive."

The second area of review at step two involves screening the efficacy studies to see if they use the type of study design required by the WWC. As described more fully in the companion report, the WWC only accepts studies that employ a pretest-posttest control group design and gives its highest ratings only to studies that assign subjects to groups through random assignment. These restrictions regarding study design contrast markedly with the standard methodological literature. The standard literature recommends a variety of designs for field settings and, in fact, explicitly notes problems in using randomized designs in organizational environments such as schools.

Of the 74 efficacy studies that were identified by the WWC, 43 were rejected from consideration at step 2 for having an unacceptable design. However, 25 of these studies (almost 60%) used comparison group designs commonly accepted within the social sciences and recommended for school settings. These studies should have, arguably, been used in a systematic review of the literature on the efficacy of RM.

## Errors at Step Three

The third step of the WWC review process is designed to review studies that have an "acceptable" research design to see if they conform to WWC standards. These standards involve regulations regarding the magnitude of differences between comparison groups at pretest, the number of units (classrooms, schools, and districts) in each comparison group, and issues regarding confounds to the design. The WWC rejected all of the remaining 31 efficacy studies at this stage. However, our analysis determined that, if standard

methodological criteria had been used, almost half of these studies (n=15), should have been included in a review.

In total, of the 74 efficacy studies that the WWC identified as fitting the review protocol for Beginning Reading, our analysis indicated that over half (n=40 or 54%) should have been examined. Each of these studies employed a comparison group design that is commonly accepted within the methodological literature. Note that when the 42 studies not included in the WWC listing are considered, there are over 80 studies of *RM* that fit the WWC's protocol.

## The Effectiveness of Reading Mastery

We were able to compute effect sizes (Cohen's d) for 38 analyses listed by the WWC. All of these studies used a comparison group design that is recommended by the standard methodological literature for field settings such as schools.[28] Effect sizes are a common metric used to summarize the impact of educational interventions. They summarize the difference between an intervention and a control condition as a percentage of the common standard deviation (variability) of the two groups. An effect size of zero indicates no difference between the groups. Traditionally, effect sizes of .25 or larger have been seen as educationally important or educationally significant. Table 4 summarizes the distribution of average effect sizes found in the analyses. They ranged from -.53 to 2.44. Only one effect was smaller than -.25 (indicating an advantage for the comparison program). In contrast, three-fourths of the effects were larger than the .25 criterion, and the average effect across all the comparisons was .57, more than twice the .25 level. This average value is similar to the average effect sizes found in meta-analyses conducted by other researchers cited earlier in this report.

Table 4
*Effect Sizes in Studies of Reading Mastery with Beginning Readers Rejected for Inclusion in the WWC 2014 Report, Descriptive Statistics*

| | |
|---|---|
| Average | 0.57 |
| Minimum | -0.53 |
| Maximum | 2.44 |
| Number with Effect < = -.25 | 1 |
| Number with Effect > = .25 | 28 |

[28] This analysis treated the two designs included in O'Brien and Ware (2002) separately, paralleling the listings in WWC (2013b). The mixed model statistical analysis described in the companion report was conducted in two ways. One used the designs as the level two variable, essentially looking at differences by design. The other combined studies across sites, using site of a study as the level two variable. Two studies were omitted from the computation of effect sizes. One was a norm comparison design for which we were unable to obtain norms by the time this report was written (Watkins, 2008). The other was a single subject design. Results of these studies were also positive.

## Needed Changes in WWC Procedures and Policies

The analyses in this and the companion report indicate that, if the WWC is to meet its goal of providing accurate and reliable information on what works in education, there must be substantial changes in WWC procedures and policies.

The discussion above documented a number of areas in which the WWC reviewers erred in their analysis of studies of *Reading Mastery* ranging from an incomplete search of the literature, errors in the listing of relevant articles, not consulting bibliographies of literature reviewed or examining references sent them, to mistaken interpretations of the articles that were reviewed. The published descriptions of the WWC procedures do not provide the details needed to understand why these errors occurred. But, at the least, the results indicate the need to improve standards of review and that the WWC should adopt procedures that more closely mirror standard methodological practices. This would include multiple independent checks of analyses by WWC staff, having reviewers trained in relevant substantive and methodological issues, careful and thorough inspection of all analyses by independent outside reviewers, careful comparison of results to the scholarly literature, and full acknowledgement of any discrepancies with other analyses.

Even if the WWC review procedures were error free, the serious problems with WWC policies would very likely result in inaccurate reports. As can be seen in the discussion above and as described much more thoroughly in the companion report, many of the errors reported here involve WWC policies that differ dramatically from long established and well tested procedures of the scientific community. To date, the WWC has not provided scholarly justification for their dramatic departure from the standard procedures. Just as important, the statistical results in the companion technical report (as well as in Stockard, 2013b) indicate that application of the policies does not provide more precise estimates of a program's impact. Instead, by providing a very limited picture of the available evidence the reports can be very highly biased. Thus, a variety of policy changes would seem advisable such as accepting the full range of research designs recommended for field settings, including schools; altering the standards regarding equivalence of groups at pretest; and applying the "one unit" rule only when it actually results in "confounded" results.

Most important, the WWC should always compare its conclusions with the extant scholarly literature, a regular step included in reviews that conform to standard methodological procedures. When differences occur it is incumbent upon the researcher to understand why they occur and to explain the discrepancies in a report. However, as explained in correspondence with the NIFDI Office of Research, the WWC does not compare its results with those found by others.[29] Clearly, however, engaging in such comparisons could

[29] The WWC procedures state that reviews should identify relevant existing systematic reviews and meta-analyses to ensure that we have identified all of the relevant literature. However, in response to a specific query regarding whether "the WWC reviews these meta-analyses while conducting their reviews and whether

significantly increase the probability that the WWC reports would more accurately reflect the extant literature.

Since 2008 NIFDI's Office of Research and Evaluation has made numerous attempts to suggest changes to WWC policies and procedures that could enhance the accuracy of their reports. To date these efforts have had relatively little impact. Some of the reported errors have been corrected, but most remain.[30] In addition, there has been no change in the highly flawed policies.[31] Given the very severe problems documented in this and the companion report, users should be very wary of the conclusions presented in WWC reports.[32] Those seeking reliable information on the efficacy of educational programs would be well advised to consult the standard research literature instead.

---

they consider including the differing opinions on the effectiveness of the programs reviewed in the reports," the WWC responded in an e-mail on February 28, 2014, "The answer is no. As stated previously, other meta-analyses may differ in their inclusion criteria and standards. We do not report on or interpret the findings from other such reviews. We do list them in our citations so interested readers may find them." In other words, the WWC reviews meta-analyses to identify studies, but, contrary to standard practice in the field, does not compare their results to those within the scholarly community.

[30] A 2008 report of the WWC on *Reading Mastery* for beginning readers included numerous errors, which were reported to the WWC. However, as documented in Stockard (2008), the WWC failed to correct any of these mistakes. The WWC review of the use of RM with students with Learning Disabilities also had severe errors of interpretation. The WWC corrected one of those errors, but refused to correct the other in which the WWC's conclusion directly contradicted the results of the research. (Stockard & Wood, 2013). In addition, as noted earlier in the text, the WWC corrected the errors in the listing of articles in the original November, 2013, report on the use of RM for beginning readers that had been reported to them. However, the nature of the errors was not publicly acknowledged, nor was the official date of the report changed.

[31] NIFDI's attempts to provide input on policies may have, ironically, resulted in the policies becoming even more removed from standard methodological practices. See the discussion in the companion report (NIFDI Technical Report 2014-3, page 16) regarding the 2014 version of the WWC's *Procedures and Standards Handbook*, which, for the first time, included a determination that studies comparing results from cohorts within a school would not be allowed. As noted in the text, the cohort comparison design is explicitly recommended by the CCSS tradition, and NIFDI had brought this design and its recommendation to the attention of the WWC in 2011 and in other communications.

[32] A forthcoming NIFDI Technical Report (2014-5) will document the numerous errors others have reported to the WWC.

# Appendix A
# Studies Included in the WWC 2013 Report on *Reading Mastery* for Beginning Reading

(Studies marked with an asterisk(*) involve a program other than *Reading Mastery*)

Airhart, K. M. (2005). *The effectiveness of Direct Instruction in reading compared to a state mandated language arts curriculum for ninth and tenth graders with specific learning disabilities* (Unpublished doctoral dissertation). Tennessee State University, Nashville.

Algozzine, B., Wang, C., White, R., Cooke, N., Marr, M. B., Algozzine, K., Helf, S. S., & Duran, G. Z. (2012). Effects of multi-tier academic and behavior instruction on difficult-to-teach students. *Exceptional Children, 79*(1), 45-64.

Asfendis, G. (2008). *Phonemic awareness and early intervention: An evaluation of a pilot phonemic awareness program*. Dissertation Abstracts International, 62.

Ashworth, D. R. (1999). Effects of Direct Instruction and basal reading instruction programs on the reading achieve-ment of second graders. *Reading Improvement, 35*(4), 150–156.

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003a). City Springs Elementary School, Baltimore, MD. In *Results with Reading Mastery* (pp. 14–15). New York: McGraw-Hill.

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003b). Eshelman Avenue Elementary, Lomita, CA. In *Results with Reading Mastery* (pp. 16–17). New York: McGraw-Hill.

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003c). Fort Worth Independent School District, Fort Worth, TX. In *Results with Reading Mastery* (pp. 4–5). New York: McGraw-Hill.

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003d). Lebanon School District, Lebanon, PA. In *Results with Reading Mastery* (pp. 8–9). New York: McGraw-Hill

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003e). Park Forest-Chicago Heights School District 163, Chicago, IL. In *Results with Reading Mastery* (pp. 10–11). New York: McGraw-Hill.

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003f). Portland Elementary School, Portland, OR. In *Results with Reading Mastery* (pp. 2–3). New York: McGraw-Hill.

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003g). Wilson Primary School, Phoenix, AZ. In *Results with Reading Mastery* (pp. 6–7). New York: McGraw-Hill.

Association for Supervision and Curriculum Development and the Council of Chief State School Officers. (2003h). Roland Park Elementary/Middle School, Baltimore, MD. In *Results with Reading Mastery* (pp. 12–13). New York: McGraw-Hill.

Batchelder, H. L. W. (2008). *An investigation of the efficacy of the text talk strategy on pre-school students' vocabulary acquisition*. Retrieved from http://purl.fcla.edu

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

Borges, J. (2009). *Reciprocal teaching strategies in context: Implications for sixth grade humanities*. New York: Bank Street College of Education.

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

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

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.

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

*Cohen, E. J., & Brady, M. P. (2011). Acquisition and generalization of word decoding in students with reading disabilities by integrating vowel pattern analysis and children's literature. *Education and Treatment of Children, 34*(1), 81–113.

Collier, P. R. (2008). *The impact of literacy coaching on teacher fidelity and students with learning disabilities' reading achievement*. Dissertation Abstracts International, 70 (02A), 126–514.

Comprehensive School Reform Quality Center. (2006). *CSRQ center report on elementary school CSR models*. Washington, DC: Comprehensive School Reform Quality Center, American Institutes for Research.

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

Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology, 47*, 187–214.

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

De La Cruz, C. F. (2009). *A program evaluation study of a literacy initiative for students with moderate to severe disabilities*. Retrieved from http://purl.fcla.edu

Elias, E. I. (2009). *The lived experiences of six first-grade teachers using Reading Mastery Plus curriculum in high poverty schools*. Dissertation Abstracts International, 70(7A), 182–2440.

Eppley, K. (2011). *Reading Mastery* as pedagogy of erasure. *Journal of Research in Rural Education, 26*, 1–5.

*European Centre for Reading Recovery. (2012). *Reading Recovery annual report for the UK and the Republic of Ireland: 2011-12*. University of London: Institute of Education.

*Flores, M. M., & Ganz, J. B. (2007). Effectiveness of Direct Instruction for teaching statement inference, use of facts, and analogies to students with developmental disabilities and reading delays. *Focus on Autism & Other Developmental Disabilities, 22*(4), 244–251.

Foorman, B. R., Chen, D. T., Carlson, C., Moats, L., Francis, D. J., & Fletcher, J. M. (2003). The necessity of the alphabetic principle to phonemic awareness instruction. *Reading and Writing, 16*(4), 289–324.

Foorman, B. R., Fletcher, J. M., & Francis, D. J. (2004). Early reading assessment. In W. M. Evers & H. J. Walberg (Eds.), *Testing student learning, evaluating teaching effectiveness* (pp. 81–125). Stanford, CA: Hoover Institution Press.

Foorman, B. R., Schatschneider, C., Eakin, M. N., Fletcher, J. M., Moats, L. C., & Francis, D. J. (2006). The impact of instructional practices in grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology, 31*(1), 1–29.

Foorman, B., & Al Otaiba, S. (2009). How children learn to read: Current issues and new directions in the integration of cognition, neurobiology and genetics of reading and dyslexia research and practice. In K. Pugh & P. McCardle (Eds.), *Reading remediation: State of the art* (pp. 257–274). New York: Psychology Press.

Fredrick, L. D., Keel, M. C., & Neel, J. H. (2002). Making the most of instructional time: Teaching reading at an accelerated rate to students at risk. *Journal of Direct Instruction, 2*(1), 57–63.

Frijters, J. C., Lovett, M. W., Steinbach, K. A., Wolf, M., Sevcik, R. A., & Morris, R. D. (2011). Neurocognitive predictors of reading outcomes for children with reading disabilities. *Journal of Learning Disabilities, 44*(2), 150–166.

Fulmer, S. M., & Frijters, J. C. (2011). Motivation during an excessively challenging reading task: The buffering role of relative topic interest. *Journal of Experimental Education, 79*(2), 185–208.

Goss, C. L., & Brown-Chidsey, R. (2012). Tier 2 reading interventions: Comparison of Reading Mastery and Fundations Double Dose. *Preventing School Failure, 56*(1), 65–74.

Graves, A. W., Duesbery, L., Pyle, N. B., Brandon, R. R., & McIntosh, A. S. (2011). Two studies of Tier II literacy development: Throwing sixth graders a lifeline. *The Elementary School Journal, 111*(4), 641–661.

Greaney, K., & Arrow, A. (2012). Phonological-based assessment and teaching within a first year reading program in New Zealand. *Australian Journal of Language & Literacy, 35*(1), 9–32.

Green, A. K. (2010). *Comparing the efficacy of SRA Reading Mastery and guided reading on reading achievement in struggling readers*. Dissertation Abstracts International, 71(11A), 3969.

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

*Harvey, M. W. (2012). *Union County Public Schools action research: Comparing early literacy interventions used in Union County Public Schools; Reading Recovery vs. Leveled Literacy Intervention*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 73(3A), 989.

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

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

*Hudler, S. E. (2008). *A description of explicit phonological instruction for elementary children with mild disabilities with computer-assisted instruction*. Columbus: Ohio State University.

Humphries, T., Neufeld, M., Johnson, C., Engels, K., & McKay, R. (2005). A pilot study of the effect of Direct Instruction programming on the academic performance of students with intractable epilepsy. *Epilepsy & Behavior, 6*(3), 405–412.

Intensive, tailored tuition raise literacy. (2012). *Children and Young People Now, 2*(15), 32–33.

Jones, C. D. (2002). *Effects of Direct Instruction programs on the phonemic awareness abilities of kindergarten students*. Dissertation Abstracts International, 63(03), 902A.

Jordan, N. L. (2005). Basal readers and reading as socialization: What are children learning? *Language Arts, 82*(3), 204–213.

Joseph, B. L. (2000). *Teacher expectations of low-SES preschool and elementary children: Implications of a research-validated instructional intervention for curriculum policy and school reform*. Dissertation Abstracts International, 65(01), 35A.

Joseph, L. M., & Schisler, R. (2009). Should adolescents go back to the basics? A review of teaching word reading skills to middle and high school students. *Remedial and Special Education, 30*(3), 131–147.

Kaiser, S., Palumbo, K., Bialozor, R. C., & McLaughlin, T. F. (1989). The effects of Direct Instruction with rural remedial education students: A brief report. *Reading Improvement, 26*(1), 88–93.

Kamps, D. M., & Greenwood, C. R. (2005). Formulating secondary-level reading interventions. *Journal of Learning Disabilities, 38*(6), 500–509.

Kamps, D., Abbott, M., Greenwood, C., Arreaga-Mayer, C., Wills, H., Longstaff, J.,...Walton, C. (2007). Use of evidence-based, small-group reading instruction for English language learners in elementary grades: Secondary-tier intervention. *Learning Disability Quarterly, 30*(3), 153–168.

Kamps, D., Abbott, M., Greenwood, C., Wills, H., Veerkamp, M., & Kaufman, J. (2008). Effects of small-group reading instruction and curriculum differences for students most at risk in kindergarten. *Journal of Learning Disabilities, 41*(2), 101–114.

Kanfush, P. M., III. (2010). *Use of Direct Instruction to teach reading to students with significant cognitive impairments: Student outcomes and teacher perceptions*. Dissertation Abstracts International, 71(12A), 4355.

*Keafer, K. A. (2008). *Effects of National Institute for Learning Development educational therapy for students with learning difficulties*. Dissertation Abstracts International, 69(06A), 123-222.

Kubina, R. M., Jr., Commons, M. L., & Heckard, B. (2009). Using precision teaching with Direct Instruction in a summer school program. *Journal of Direct Instruction, 9*(1), 1–12.

Kuder, S. J. (1990). Effectiveness of the DISTAR reading program for children with learning disabilities. *Journal of Learning Disabilities, 23*(1).

Kuder, S. J. (1991). Language abilities and progress in a Direct Instruction reading program for students with learning disabilities. *Journal of Learning Disabilities, 24*(2).

*Laska, K. (2009). *Computer-based phonological awareness training for adolescent English language learners*. Retrieved from http://www.hamline.edu

League, M. B. (2001). *The effect of the intensity of phonological awareness instruction on the acquisition of literacy skills*. Dissertation Abstracts International, 62(10), 3299A.

LeClair, C. M. (2011). *Determining the longitudinal effects of acculturation orientation on elementary-aged Spanish-speaking English language learner students' reading progress* (Unpublished doctoral dissertation). University of Nebraska, Lincoln.

Lehr, F., & Osborn, J. (2012). *Reading, language, and literacy*. Retrieved from http://www.eblib.com

*Lewis, D. (2012). *The effects of Reading Recovery on the English reading development of native Spanish-speaking ESOL students*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 72(9A), 3200.

Lovett, M. W., Lacerenza, L., De Palma, M., & Frijters, J. C. (2012). Evaluating the efficacy of remediation for struggling readers in high school. *Journal of Learning Disabilities, 45*(2), 151–169.

Mac Iver, M. A., & Kemper, E. (2002). The impact of Direct Instruction on elementary students' reading achievement in an urban school district. *Journal of Education for Students Placed at Risk, 7*(2), 197–220.

Marchand-Martella, N. E., Martella, R. C., Kolts, R. L., Mitchell, D., & Mitchell, C. (2006). Effects of a three-tier strategic model of intensifying instruction using a research-based core reading program in grades K–3. *Journal of Direct Instruction, 6*(1), 49–72.

Marchand-Martella, N. E., Ruby, S. F., & Martella, R. C. (2007). Intensifying reading instruction for students within a three-tier model: Standard-protocol and problem

solving approaches within a response-to-intervention (RTI) system. *TEACHING Exceptional Children Plus, 3*(5).

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

Massar, E. M., Widener University, & School of Human Service Professions. (2009). *A case study using the Corrective Reading program in a junior/senior high remedial class* (Unpublished doctoral dissertation). Widener University, Chester, PA.

Mathes, P. G., & Proctor, T. J. (1988). Direct Instruction for teaching "hard to teach" students. *Reading Improvement, 25*(2), 92–97.

*McClendon, I. D. (2012). *A longitudinal case study of a literacy program titled Reading Recovery for students in a struggling Midwestern school district*. Dissertation Abstracts International Section A: Humanities and Social Sciences, 73(4A), 1357.

McCollum, S., McNeese, M. N., Styron, R., & Lee, D. E. (2007). A school district comparison of reading achievement based on three reading programs. *Journal of At-Risk Issues, 13*(1), 1–6.

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

Nanda, A. O., & Fredrick, L. D. (2007). The effects of combining repeated reading with "Reading Mastery" on first graders' oral reading fluency. *Journal of Direct Instruction, 7*(1), 17–27.

Neely, M. (1995). The multiple effects of whole language, precision teaching and Direct Instruction on first-grade story-reading. *Effective School Practices, 14*(4), 33–42.

O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth independent school district. *Journal of Education for Students Placed at Risk, 7*(2), 167–195.

O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth independent school district. *Journal of Education for Students Placed at Risk, 7*(2), 167–195.

O'Connor, R. E., Harty, K. R., & Fulmer, D. (2005). Tiers of intervention in kindergarten through third grade. *Journal of Learning Disabilities, 38*(6), 532–538.

Palacios, N. (2009). *Immigration, child development, and early education in the twenty-first century*. Dissertation Abstracts International, 70(12A), 201-4568.

Prager, A. J. (2008). *A comparison of linear versus spiral multiple exemplar instruction on derived abstracted textual responses of preschool children* (Unpublished doctoral dissertation). Columbia University, New York.

Ralston, N. C., Benner, G. J., Nelson, J. R., & Caniglia, C. (2009). The effects of the "language arts" strand of the "*Reading Mastery* signature series" on the reading and language skills of English language learners. *Journal of Direct Instruction, 9*(1), 47–55.

*Redding, L. R. (2012). *An investigation of the sustained effects of Reading Recovery on economically disadvantaged fifth grade students* (Unpublished doctoral dissertation). Widener University, Chester, PA.

Riepl, J. H., Marchand-Martella, N., & Martella, R. C. (2008). The effects of "Reading Mastery Plus" on the beginning reading skills of students with intellectual and developmental disabilities. *Journal of Direct Instruction, 8*(1), 29–39.

Ryder, R. J., Burton, J. L., & Silberg, A. (2006). Longitudinal study of Direct Instruction effects from first through third grades. *Journal of Educational Research, 99*(3), 180–191.

Ryder, R. J., Sekulski, J. L., & Silberg, A. (2003). *Results of Direct Instruction reading program evaluation longitudinal results: First through third grade 2000–2003*. Retrieved from: http://www.uwm.edu

*Schelling, C. M. (2010). *The development of early reading skills in a sample of kindergarten students*. Dissertation Abstracts International, 70(10B), 6578.

Schieffer, C., Marchand-Martella, N., Martella, R., & Simonsen, F. (n.d.). *The research base for Reading Mastery*. DeSoto, TX: The McGraw-Hill Companies.

Shelton, N. R. (2010). Program fidelity in two "Reading Mastery" classrooms: A view from the inside. *Literacy Research and Instruction, 49*(4), 315–333.

Shippen, M. E., Houchins, D. E., Calhoon, M. B., Furlow, C. F., & Sartor, D. L. (2006). The effects of comprehensive school reform models in reading for urban middle school students with disabilities. *Remedial and Special Education, 27*(6), 322–328.

Simmons, D. C., Coyne, M. D., Kwok, O., McDonagh, S., Harn, B. A., & Kame'enui, E. J. (2008). Indexing response to intervention. *Journal of Learning Disabilities, 41*(2), 158–173.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391–1466.

Slavin, R. E., Lake, C., Cheung, A., & Davis, S. (2009). *Beyond the basics: Effective reading programs for the upper elementary grades*. Baltimore, MD: Center for Data-Driven Reform in Education, Johns Hopkins University.

Smolkowski, K., Biglan, A., Barrera, M., Taylor, T., Black, C., & Blair, J. (2005). Schools and Homes in Partnership (SHIP): Long-term effects of a preventive intervention focused on social behavior and reading skill in early elementary school. *Prevention Science: The Official Journal of the Society for Prevention Research, 6*(2), 113–125.

SRA/McGraw-Hill. (2005a) *Test scores transform troubled school into national model*. Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (2005b). *All grade 3 students in two Monroe, Wisconsin elementary schools score proficient or advanced in reading*. Retrieved from http://www.mheresearch.com

SRA/McGraw-Hill. (2005c). *Barren County elementary schools post highest reading scores ever*. Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (2005d). *California blue ribbon school closes achievement gap with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005e). *Delaware charter school students maintain high reading scores.* Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005f). *Direct Instruction helps Kentucky blue ribbon school attain record reading scores*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005g). *Florida elementary students master reading in preparation for junior high*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005h). *Florida school raises reading scores with Direct Instruction.* Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005i). *Miami elementary school boosts FCAT scores with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.
*SRA/McGraw-Hill. (2005j). *Milwaukee elementary nearly doubles reading scores*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005k). *Oregon Reading First project uses Reading Mastery Plus as core reading program*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005l). *Phoenix inner-city students strive toward national reading average.* Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005m). *Reading Mastery helps Florida students advance two grade levels in reading*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005n). *Reading Mastery Plus helps Colorado school achieve AYP for first time*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2005o). *Washington elementary students excel on WASL, ITBS with Reading Mastery Plus*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006a). *Cleveland school keeps Reading Mastery as curriculum core.* Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006b). *DIBELS scores advance to grade level with Reading Mastery.* Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006c). *Exceptional education and regular education students excel with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006d). *Florida school moves from D grade to A with Reading Mastery.* Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006e). *Native American school uses Reading First grant to implement Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006f). *Reading Mastery, Corrective Reading help students with disabilities achieve significant academic growth*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006g). *Reading proficiency more than doubles among Putnam County special education students*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2006h). *Struggling Milwaukee readers make strong gains with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (2006i). *Utah school district maintains high language arts scores with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2007a). *Direct Instruction reduces special education referrals in Louisiana school district by half*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2007b). *Low-performing Kentucky school on its way to high-performing with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2007c). *Reading Mastery helps special education students meet state reading standards*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2007d). *Reading scores rise at Alabama elementary school with Reading Mastery Plus*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2007e). *SRA/McGraw-Hill's reading programs bring increases in Baltimore's scores*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2007f). *Title I schools in North Carolina district meet all-state reading targets with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (2009a). *A report on the effects of SRA/McGraw-Hill's Reading Mastery, Signature Edition: A response to intervention solution*. DeSoto, TX: Author.
SRA/McGraw-Hill. (n.d.a) *Seattle school boosts reading scores with Reading Mastery curriculum*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.b). *Anchorage school's diverse population flourishes with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.c). *Combination of Open Court reading and Direct Instruction equal consistently high reading scores*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.d). *Commitment to Direct Instruction increases reading scores at Cleveland school*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.e). *Direct Instruction drives success for bilingual students at Houston elementary school*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.f). *ELL and struggling students at Wisconsin district build literacy skills with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.g). *High percentage of charter school's students testing above national average*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.h). *Houston blue ribbon school's deaf population achieves AYP with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.i). *Modesto elementary school advances from underperforming to distinguished with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.j). *Nebraska Reading First school reaches states highest scores with SRA reading programs*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.k). *Open Court reading and Reading Mastery combine to create successful elementary reading program*. Columbus, OH: The McGraw-Hill Companies.
SRA/McGraw-Hill. (n.d.l). *Pennsylvania SD uses Horizons to give at-risk readers foundations for success*. Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (n.d.m). *Phoenix inner-city students strive toward national reading average.* Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (n.d.n). *Results with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (n.d.o). *Special education students at California elementary school achieve AYP with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (n.d.p). *Success begins early at Alaskan elementary school*. Columbus, OH: The McGraw-Hill Companies.

SRA/McGraw-Hill. (n.d.q). *Wisconsin teachers use Horizons to customize reading lessons.* Columbus, OH: The McGraw-Hill Companies.

Stockard, J. (2008). *The long-term impact of NIFDI-supported implementation of direct instruction on reading achievement: An analysis of fifth graders in the Baltimore City Public School System*. Eugene, OR: National Institute for Direct Instruction.

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

Stockard, J. (2010b). *Merging the accountability and scientific research requirements of the No Child Left Behind act: Using cohort control groups*. Eugene, OR: National Institute for Direct Instruction.

Stockard, J. (2010c). *The impact of Reading Mastery in kindergarten on reading achievement through the primary grades: A cohort control group design*. Eugene, OR: National Institute for Direct Instruction.

Stockard, J. (2011). Increasing reading skills in rural areas: An analysis of three school districts. *Journal of Research in Rural Education, 26*(8).

Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's Reading Mastery. *Journal of Behavior Assessment & Intervention in Children, 1*(1), 2–24. Study A.

Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's Reading Mastery. *Journal of Behavior Assessment & Intervention in Children, 1*(1), 2–24. Study B.

Sturtz, T. I. (2009). *Preservice literacy instruction and the benefits of Direct Instruction.* Charleston, SC: BiblioBazaar.

Thames, D., Kazelskis, R., & Kazelskis, C. R. (2006, November). *Reading performance of elementary students: Results of a five-year longitudinal study of direct reading instruction*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Birmingham, AL.

Thomson, B. (1991). Pilot study of the effectiveness of a Direct Instruction model (Reading Mastery Fast Cycle) as a supplement to a literature based delivery model (Houghton-Mifflin Integrated Reading Program) in two regular first grade classrooms. *Florida Educational Research Council Research Bulletin, 23*(2), 3–23.

Torgesen, J. K. (2002). Lessons learned from intervention research in reading: A way to go before we rest. *Learning and Teaching Reading,* 89–103.

Trout, A. L., Epstein, M. H., Mickelson, W. T., Nelson, J. R., & Lewis, L. M. (2003). Effects of a reading intervention for kindergarten students at risk for emotional disturbance and reading deficits. *Behavioral Disorders, 28*(3), 313–326.

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

Van Norman, R. K., & Wood, C. L. (2008). Effects of prerecorded sight words on the accuracy of tutor feedback. *Remedial and Special Education, 29*(2), 96–107.

Viel-Ruma, K. A. (2008). *The effects of Direct Instruction in writing on English speakers and English language learners with disabilities*. Dissertation Abstracts International, 69(07A), 149-2672.

Watkins, T. B. (2008). *A comparative analysis of the effectiveness of Direct Instruction reading on African American, Caucasian, and Hispanic students*. Dissertation Abstracts International, 69(03A), 104-923.

*Welsh, D. K. (2010). *Effects of differentiated instruction and word attack strategies on struggling readers*. Retrieved from http://www.proquest.com

Wills, H., Kamps, D., Abbott, M., Bannister, H., & Kaufman, J. (2010). Classroom observations and effects of reading interventions for students at risk for emotional and behavioral disorders. *Behavioral Disorders, 35*(2), 103–119.

Wilson, P., Martens, P., & Arya, P. (2005). Accountability for reading and readers: What the numbers don't tell. *Reading Teacher, 58*(7), 622–631.

Wiltz, N., & Wilson, G. P. (2006). An inquiry into children's reading in one urban school using SRA Reading Mastery (Direct Instruction). *Journal of Literacy Research, 37*(4), 493–528.

Wyse, D. (2012). *Literacy teaching and learning*. London: SAGE Publications.

Zayac, R. M. (2008). *Direct Instruction reading: Effects of the Reading Mastery Plus-Level K curriculum on pre-school children with developmental delays*. Dissertation Abstracts International, 69(10B), 226-6458.

# Appendix B
# Studies of *Reading Mastery* Omitted from the WWC Report

To develop this list of studies regarding Reading Mastery that could have been included in the review the WWC protocol regarding acceptable studies was followed as closely as possible. Only studies published in 1983 or later were included. Studies that appeared to be about students with disabilities or special education were excluded as were those in which the students began their study with *RM* in grades 4 or higher. All of the works provide data sufficient to calculate effect sizes and will be used in a forthcoming meta-analysis. Items marked with an asterisk were included in NIFDI's response to the WWC's 2008 report on *Reading Mastery* (Stockard, 2008, pp. 43-47). Thus the WWC had been informed of their relevance.

Articles marked by (a) at the end of the reference were cited by 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, which is in the WWC list of articles. Those marked with (b) were cited by Comprehensive School Reform Quality Center. 2006. *CSRQ Center Report on Elementary School Comprehensive School Reform Models*. Washington, D.C. American Institutes for Research, also in the WWC listing. It appears that the WWC did not examine the reference list of these articles, even though such a perusal is a standard part of a literature review.

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

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

Bressi, T., Bressi, R., Engelmann, K., Johnston, A., Silbert, J., & Stockard, J. (2010). Direct Instruction in Africa, *DI News, Summer, 10* (2), 6-8.

Brooks-Hodridge, D. (1995). *Effects of Interactive Story Reading on Concepts about Print and Journal Writing in First-Grade Children,* Unpublished D.Ed. Dissertation, Texas Women's University, Denton Texas.

Brumbley, S. A. (1998). *The effects of a first grade phonological awareness intervention in reducing special education referrals*. Unpublished Masters thesis, University of Oregon, Eugene, OR.

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

Cross, R. W., Rebarber, T., & Wilson, S. F. (2002). Student gains in a privately managed network of charter schools using Direct Instruction. *Journal of Direct Instruction, 2*(1), 3-21.

Darch, C., Gersten, R., Taylor, R. (1987). Evaluation of the Williamsburg County Direct Instruction program: Factors leading to success in rural elementary programs. *Research in Rural Education, 4*(3), 111-118.

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

Engelmann, S., & Carnine, D. (1982). DI outcomes with middle-class second graders. *ADI News, 1*(2), 2-5, reprinted *ADI News*, Winter, 1989, pp. 2-5.

Francis, B.J. (1991) *Matching reading programs to students' needs: An examination of alternate programming using a direct instruction program in the regular classroom* (Master's thesis, Simon Fraser University). Masters Abstracts International, 31(01), 144-61.

Frink-Lawrence, V. (2003). *Closing the achievement gap: The implementation of Direct Instruction in Whiteville City Schools*. MS Thesis, Watson School of Education, University of North Carolina at Wilmington. http://libres.uncg.edu/ir/uncw/f/frink-lawrencev2003-1.pdf.

Gersten, R. (1985). Structured immersion for language minority students: Results of a longitudinal evaluation. *Educational Evaluation and Policy Analysis, 7* (3), 187-196 (republished in 1997 in *Effective School Practices, 16*(3), 21-29.

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

*Gersten, R. M. Darch, C., & Gleason, M. (1988). Effectiveness of a Direct Instruction academic Kindergarten for low-income students. *The Elementary School Journal, 89*(2), 227-240.

*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)

Goldman, B. (2000). *A study of the implementation of a Direct Instruction reading program and its effects on the reading achievement of low socioeconomic students in an urban public school*. Unpublished doctoral dissertation, Loyola University, Chicago IL.

Grossen, B. & Kelly, B.F. (1992b). Using Direct Instruction improve the effectiveness of teachers in South Africa. *South African Journal of Education, 12*, 143-147.

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

Herrington, M.S. (1999). *A Comparative Analysis of the Reading Mastery and Sliver Burdett Reading Programs for Elementary Students*. Unpublished D.Ed. Dissertation, Mississippi State University.

Johnson, S. (1985). *The effects of using the Reading Mastery Direct Instruction program with average and above-average kindergarteners: A pilot study*. Unpublished master's thesis, University of Washington, Seattle, WA.

*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.

*Kamps, D., Wills, H., Greenwood, C., Thorne, S. Lazo, J., et al. (2003). Curriculum influences on growth in early reading fluency for students with academic and behavioral risks. *Journal of emotional and behavioral disorders, 11*(4), 211-224.

*Keel, M. C., Fredrick, L. D., Hughes, T. A., & Owens, S. H. (1999). Using paraprofessionals to deliver Direct Instruction reading programs. *Effective School Practices 18*(2), 16-22.

*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.

*Ligas, M. R. (2002). Evaluation of Broward County Alliance of Quality Schools Project. *Journal of Education for Students Placed At Risk, 7*(2), 117-139.

Lovett, M. W., Palma, M., Frijters, J., Steinbach, K., Temple, M., Benson, N., & Lacerenza, L. (2008). Interventions for reading difficulties: A comparison of response to intervention by ELL and EFL struggling readers. *Journal of Learning Disabilities, 41*, (333-352).

McGahey, J. (2002). *Differences between a Direct Instruction reading approach and a balanced reading approach among elementary school students*. Dissertation Abstracts International, 63 (06A), 2147. (UMI No. 3057184)

Meyer, L.A. (1984). Long-term academic effects of the direct instruction project follow through. *The Elementary School Journal, 84*(4), 380-394.

Ocokoljich, E.D. (1997). *The effects of Reading Mastery I and II on the reading achievement of first and second grade students identified as having low phonological awareness skills*. Unpublished master's thesis. University of Wisconsin-Madison, Madison, WI.

Pechous, D. J. (2012). *Minimizing Reading Regression through a Direct Instruction Summer Reading Program*. Unpublished Doctoral Dissertation, University of Nebraska, Lincoln.

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

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

*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. (b)

*Sexton, C. W. (1989). Effectiveness of the DISTAR Reading I program in developing first graders' language skills . *Journal of Educational Research, 82*(5), 289-293.

Slocum, T. (2000). *Brief report on performance of students whose first language is Spanish: Direct Instruction and a comparison school*. Unpublished paper, Utah State University

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

Stockard, J. (2008). *Reading Achievement in a Direct Instruction School and a "Three Tier" Curriculum School*. National Institute for Direct Instruction, Technical Report 2008-5.

Stockard, J. (2011a). Direct Instruction and first grade reading achievement: The role of technical support and time of implementation. *Journal of Direct Instruction, 11* (1), 31-50.

Stockard, J. (2011b). *Changes in reading achievement at a Florida Elementary School: A randomized control study of Reading Mastery*, NIFDI Technical Report.

Sullivan, M. (2002). *Reading Mastery versus word study instruction as it pertains to third graders' reading achievement scores*. Unpublished eduational specialist's thesis. Western Kentucky University, Bowling Green, KY

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

*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)

Vitale, M. & Joseph, B. (2008). Broadening the institutional value of Direct Instruction Implemented in a low-SES elementary school: Implications for scale-up and school reform. *Journal of Direct Instruction, 8*(1), 1-18

Wrobel, S. (1996). *The effectiveness of Direct Instruction on the various reading achievement categories.* Technical Report.ERIC report, ED395292.

*Yu, L. and Rachor, R. (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, LA.(ED 441831)

# Appendix C
# Studies Rejected at Step 2 that Could Have Been Reviewed

As described in the text, our review identified 25 efficacy studies of RM that were rejected from consideration at step 2 because of issues related to the study's design but that, using standard methodological criteria, could have been considered for review. All of the studies involved comparison groups. Summaries of these 25 studies are below including descriptions of the studies' designs, possible objections that the WWC might have to each study, and calculated effect sizes. The numbers associated with each article are referenced in tables included in Appendix E. The reason that the WWC gave for rejecting each article is given in parentheses.

1) Butler, P. A. (2003). Achievement outcomes in Baltimore City Schools. *Journal of Education for Students Placed at Risk, 8*, 33–60. (rejected by WWC for unacceptable design)

This article reports changes in student achievement in the Baltimore City Public School System after a "Master Plan" of improving student achievement was introduced. Two other studies have also examined data from the BCPSS over this time period, using slightly different approaches. As described below, the Mac Iver and Kemper (2002) study was rejected because "the implementation was not implemented as designed." Stockard (2011a) was not included in the list of studies reviewed by the WWC, and is one of those listed in Appendix B.

After implementation of the Master Plan, Direct Instruction was implemented in some of the BCPSS schools. Butler reports comparisons of the schools that used DI with "schools with similar achievement histories that had not implemented this approach" (pp. 43-44). Elementary schools were matched on achievement on reading scores on the 1993 state assessment for fifth graders. Schools began implementation of the program at different years and data for these cohorts are reported separately. Data were reported on the percentage of third grade students attaining satisfactory performance on the Maryland School Performance Assessment Program (p. 45) and normal curve equivalent (NCE) scores on the reading portion of the Comprehensive Test of Basic Skills for second graders (p. 47) in the two groups of schools. Effect sizes ranged from -.57 to .33, with an average of -.11.

As noted above, the WWC rejected the Butler study because it did not use an acceptable design. It might also reject the study because it does not mention *Reading Mastery* specifically, but instead describes Direct Instruction programs in general. In addition, although schools were matched on scores in a prior year, these prior scores were not reported. In addition, while the schools were matched, the post-data on the control schools

was not reported separately by school or cohort of schools, but as a total group. Finally, it could reject the study based on the "one unit" rule, for all the data came from one district.

2) Fredrick, L. D., Keel, M. C., & Neel, J. H. (2002). Making the most of instructional time: Teaching reading at an accelerated rate to students at risk. *Journal of Direct Instruction, 2*(1), 57–63. (rejected by WWC for unacceptable design)

Fredrick and associates examined "rate of reading gain" on the Woodcock Reading Mastery Test of first and second grade students when they were instructed in *Reading Mastery* rather than with a whole language approach. A pretest-posttest repeated measures design was used in which students' reading gains over the year when they had whole language instruction was compared to their gains when they had *RM*. Seventy-seven students from one school (44 in first grade and 63 in second grade) were included in the analysis. The rate of reading gain (using the Bonferroni correction to adjust for multiple tests) was significantly greater when students had *RM* than when they had whole language instruction for two of the four measures in each grade. The associated effect sizes ranged from .05 to 1.72, with an average value of .58.[33]

The WWC rejected this study because it did not have an "acceptable" design, apparently rejecting the authors' explanation that "participants were used as their own controls in a repeated-measures design that compares participants to themselves at two different points in time" (p. 59). One could argue, however, that this design has strong controls for variables related to individuals' abilities.

3) Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *Journal of Special Education, 36*(2), 69–79; Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education, 39*(2), 66–85. (rejected by WWC because design was "bundled with other components")

There is one other study written by these authors on the same data set, but not included in the WWC list. It reports on the first year results in the study and was published in the same journal as the other two studies:

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

[33] The effect sizes were calculated by converting the t-ratios in the article to effect sizes using the on-line calculator at http://easycalculation.com/statistics/effect-size-t-test.

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

This group of studies is very strong methodologically, and uses the pretest-posttest control group design with random assignment that is preferred by the WWC. However, the WWC rejected the study because it "bundled" the *RM* intervention with other interventions. Part of the intervention involved behavioral training, although the authors are firm in noting that this should not be an issue. It is unclear why the WWC did not accept the authors' conclusion (especially since the articles were all published in highly respected journals). However, the WWC might still reject the study because students received either *RM* or *CR* depending upon their grade level. *Corrective Reading* is a version of *RM* designed for older students who are behind in reading. It moves at about twice the pace of *RM* to help students catch up with their peers. The authors gave no indication that results differed across grade levels. Finally, the WWC might reject the studies because pretest data weren't included in each of the articles. However, this should not be an issue because students were randomly assigned. When data on pretest scores were given (for some comparisons in 2005, at the 2nd year maintenance follow-up) the average difference at pretest was .01. The absolute value of one of the 6 comparisons exceeded .25 (albeit with higher scores for the control group and thus a conservative difference and also, of course a pretest difference that was more than offset by the positive effect size). Thus, it is likely that the WWC could reject it for unequal groups at pretest as well.[34] Of course, the groups were randomly assigned and multivariate

[34] The 2000 and 2002 reports were included in the WWC's report regarding *RM* for ELL students. However, the studies were rejected for inclusion in the reviews for beginning reading and adolescent literacy for both *Corrective Reading* and *Reading Mastery* because they didn't fit the protocol (adolescent literary) or because of a supposed confound (See Stockard and Wood, 2013).

analyses adjusted for pretest scores; corrections accepted as a matter of course by the scholarly community.

4) Joseph, B. L. (2000). *Teacher expectations of low-SES preschool and elementary children: Implications of a research-validated instructional intervention for curriculum policy and school reform*. Dissertation Abstracts International, 65(01), 35A. (rejected by WWC for unacceptable design).
   The data from this study were published in Vitale, M. & Joseph, B. (2008). Broadening the institutional value of Direct Instruction implemented in a low-SES elementary school: Implications for scale-up and school reform. *Journal of Direct Instruction, 8*(1), 1-18. This published study was not listed in the WWC's list.

This study examined the effects of implementing a school-wide Direct Instruction reading program in one low-income elementary school in the rural southeastern United States. Programs used included *Language for Learning, Reading Mastery*, and *Corrective Reading*. A cohort control group design compared the percentage of third, fourth, and fifth grade students who passed the state reading assessment. The percentage passing grew from 24 percent before implementation of the program to 71 percent after 7 years of implementation. The associated effect sizes ranged from .09 to 1.37, with an average value of .60.

The WWC reported that it rejected the study because of the design, no doubt because it does not accept the logic of the cohort control group design. They may well also object to the lack of pretest data on the individual level and the use of three DI programs. Note that the school used the three programs as they are intended. *Language for Learning* is designed to enhance students' language abilities so that it is easier for them to learn to read. *Corrective Reading* is only used with older students, primarily beyond those included in the comparisons.

5) Marchand-Martella, N. E., Martella, R. C., Kolts, R. L., Mitchell, D., & Mitchell, C. (2006). Effects of a three-tier strategic model of intensifying instruction using a research-based core reading program in grades K–3. *Journal of Direct Instruction, 6*(1), 49–72. Another article rejected by the WWC for an unacceptable design reported on the same data set was Marchand-Martell, et al, 2006: Marchand-Martella, N. E., Ruby, S. F., & Martella, R. C. (2007). Intensifying reading instruction for students within a three-tier model: Standard-protocol and problem solving approaches within a response-to-intervention (RTI) system. *TEACHING Exceptional Children Plus, 3*(5). We have not included this second analysis in our compilation of results.

This study reported changes in reading achievement for 371 students in grades K to 3 in one elementary school during a single academic year (72 in k, 86 in 1st, 82 in 2nd, and 89 in 3rd). For students in K to grade 2 pretest scores on the Dynamic Indicators of Basic Early

Literacy Skills (DIBELS) were reported from the first week of instruction in the fall and posttest scores were given from mid-May, at the end of the school year. The measures used for each year and testing period matched those recommended by the DIBELS authors. Data for fall and spring for the third graders came from the Scholastic Reading Inventory (SRI). While the authors reported t-tests comparing scores in fall and spring, we have used a norm comparison design, comparing students' gains over time to what would be expected given national norms (obtained from Good, Wallin, Simmons, Kame'enui, & Kaminski, 2002 for DIBELS and simply using the defined values for NCE scores for the SRI data (mean = 50, s.d.=21.06)).[35] The authors reported data for "typically achieving," Title I and special education students. Given the focus of the WWC protocol only results for the typical achieving students have been included in the analysis. The effect sizes ranged from -.23 (for letter naming skills in kindergarten) to .71, for the third grade SRI scores, with an average of .13.[36]

6) O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth independent school district. *Journal of Education for Students Placed at Risk, 7*(2), 167–195. (rejected by WWC for unacceptable design)

As noted above, this study was included in two places. It was rejected at step 2 for inclusion because it employed an unacceptable design, and it was rejected at step 3 because the groups were unequal prior to intervention. This probably reflects two different data analyses that were given in the article. The first used a cohort control group design, which the WWC has automatically rejected, and those results are described here. The other part used a multivariate analysis to adjust for differences in characteristics of groups. We suspect that the multivariate analysis prompted the decision that the groups were unequal. The results with the multivariate analysis are described in Appendix D.

O'Brien and Ware reported results of an evaluation of the implementation of *Reading Mastery* and *Open Court (OC)* in the Fort Worth Independent School District, replacing a whole language program that was used in earlier years. Implementations began with the lowest performing schools in the district and principals chose the reading program that they would implement. In the first year, 32 schools participated. Eighteen principals chose to use *RM* and 14 chose to use *OC*. In the second year, 2 additional schools chose *RM* and the remaining 25 used a slightly different version of *OC*. Some schools continued to use their traditional whole language approach. The students in the traditional program were less likely

[35] The effect sizes were computed by calculating the effect size associated with the comparison of the study group's scores to the national norm at fall (pre-test) and at spring (post-test). The effect size associated with the change over time is the difference of these two effect sizes. Negative differences indicate decline relative to the national norm and positive values indicate increases relative to the national norm. (See Stockard, 2013, for details.)

[36] The lower effect size for letter naming could be expected because, unlike most other programs, *Reading Mastery* does not explicitly teach letter names until a number of other skills are well established. When the value for letter naming is omitted from the analysis, the average effect size is .22.

to be minorities or to qualify for free or reduced lunch. The students in the *OC* and *RM* programs had similar socio-demographic characteristics. Training and support was provided for both the *OC* and *RM* implementations. Implementation fidelity was monitored.

Data regarding the percentage of students passing a teacher-administered reading inventory were analyzed using a cohort control group design. Data were given separately for students in grades K-2 and for the four groups of schools (*RM*, the two forms of *OC*, and the traditional whole language schools). Results were given for the fall of 1998 before implementation of the program and in the spring of 2000, after two years of implementation. The results indicated that the *RM* schools had a stronger increase than the other schools in the percentage of students passing the assessment. The associated effect sizes ranged from 0.0 to .43, with an average of .14.[37]

In addition to objections to the cohort control group with comparison group design (where changes in cohorts of two groups across time are compared), the WWC might also object to the unequal beginning scores of the two groups. Because *RM* was more often selected by principals in the neediest schools, in spite of the matching process, the percentage of students passing the assessment at the start of the intervention was substantially lower in the *RM* schools than in the control schools (d values at pre testing ranged from .41 to .49 of a s.d. lower, with an average of .46). Of course, these differences were controlled in the analysis, and the fact that the *RM* students were initially lower provides a conservative estimate of the impact.

7) SRA/McGraw-Hill. (2005b). *Barren County elementary schools post highest reading scores ever*. Columbus, OH: The McGraw-Hill Companies.

Barren County School District, based in Glasgow, Kentucky, implemented *Reading Mastery* in five elementary schools in the 2000-2001 school year. The district serves 4100 students, is 97 percent Caucasian, and has a 42 percent free and reduced lunch rate. Data were provided on CTBS scores for grade 3 students from 1999 through 2005, two years before the implementation and 5 years afterwards, separately for the five schools. The national average on the norm-referenced CTBS is 50. Before the implementation of *RM*, in 1999 and 2000, most of the schools had scores that were close to or above the national average. However, after implementation of the program, the scores increased markedly and reached their highest point when the third graders would have been exposed to the program since kindergarten (the 2004 and 2005 cohorts). The change in CTBS scores over time was analyzed using the cohort normative control group design, looking at results separately for

[37] Three steps were involved in this calculation: 1) calculating the effect size comparing the two groups of schools before the intervention occurred, reflecting the extent to which the two groups differed at pretest; 2) calculating the effect size comparing the two groups after the intervention, reflecting the extent to which the two groups differed at posttest; and 3) calculating the difference between these two effects, reflecting the amount of change in the discrepancy between the two groups in standard deviation, or effect size, form.

the five schools, thus adding a natural replication. All of the effect sizes surpassed the usual levels of educational significance, ranging from .71 to 1.66, with an average of .91. All of the t-tests were statistically significant, and the minimum value on all confidence intervals for the effect size surpassed the usual level seen as indicating educational significance.

8) SRA/McGraw-Hill. (2005d). *Delaware charter school students maintain high reading scores*. Columbus, OH: The McGraw-Hill Companies.

East Side Charter School in Wilmington, Delaware served 160 students. Over 90 percent of the students were African American and 85 percent qualified for free or reduced lunch. The school opened in 1997 and began working with Direct Instruction programs in 1998. Data were provided on the percentage of third grade East Side students meeting or exceeding state standards from 1999, the first year for which data were available, through 2005. Comparable data were also given for the state. In 1999, only 20 percent of East Side third graders met or exceeded state reading standards, but this percentage increased markedly in later years, where at least two-thirds of the students met or exceeded standards. The effect size related to the change in East Side relative to the state equaled .79. The result was statistically significant ($p = .01$), and the lower limit of the 95 percent (one-tail) confidence interval for the effect size was well above zero. (Note that in 1999 the third graders would have been exposed to RM in third grade. By 2002 students continuing in the school would have been exposed since kindergarten. Thus the comparison is between those with full exposure to the model (since kindergarten) and minimal exposure (only in third grade). The WWC could object to the inclusion of this study because no group within the Eastside data had no exposure to the program. When this study is excluded from our summary quantitative analysis the results regarding the impact of *RM* or the lack of impact of WWC criteria are unchanged.

9) SRA/McGraw-Hill. (2005e). *Direct Instruction helps Kentucky blue ribbon school attain record reading scores*. Columbus, OH: The McGraw-Hill Companies.

This Kentucky elementary school has a low income population (72% receiving free or reduced lunch) and a substantial number of ELL students (23%). The school had used *Reading Mastery* for all students in grades K-3 since 2003-2004. Data on the percentage of grade 4 students scoring at the proficient or distinguished level in reading were given for the state, the district, and the school for the year before implementation to the end of the second year of implementation. The percentage scoring at this level for the school rose from 70% to 85%, in the state the increase was from 62% to 68%, and for the district the percentage increased from 69% to 74%. Because comparative data were given, the cohort control group with historical comparison design was used. The effect size for the comparison to the state was .20, with an associated probability of .07; the effect size associated with the comparison to the district was .23, with an associated probability of .04.

10)SRA/McGraw-Hill. (2005h). *Milwaukee elementary nearly doubles reading scores*. Columbus, OH: The McGraw-Hill Companies.

Honey Creek Continuous Progress Elementary is a Pre-K to fifth grade school with 374 students in Milwaukee, Wisconsin. Forty-four percent of students received free or reduced lunch. Slightly less than ten percent were African American, fifteen percent were Hispanic, and 73 percent were Caucasian. The school began using Direct Instruction programs in the fall of 1998. The report provided data on the percentage of Grade 4 students who scored proficient or advanced on the Wisconsin Knowledge and Concepts Exam in 1997, before implementation, and from 1999 to 2005, after implementation. Before the implementation and in the first years, the fourth grade students were less likely to meet proficiency than others in the state, but this changed in later years, beginning with the cohort that had experienced *RM* since first grade. The effect size associated with the change from 1997 to 2005 was .56 and the associated t-test was statistically significant.

11)SRA/McGraw-Hill. (2005i). *Oregon Reading First project uses Reading Mastery Plus as core reading program*. Columbus, OH: The McGraw-Hill Companies.

The district in this report served slightly more than 2,000 students in grades K-12. About half of the students were Hispanic and almost all of the rest were Caucasian. Almost three-fourths (72%) of the students received free or reduced price lunch. The district adopted *Reading Mastery* as its core reading curriculum in the fall of the 2003-2004 school year. The SRA report gives the percentage of third grade students who met or exceeded state reading standards in the spring testing in 2003, the year before the program was implemented, through 2006, after three years of implementation. Data are presented for both the total group of students and for Hispanic students in the school district as well as in the state as a whole. Before *RM* was implemented the Hispanic students in the district met or exceeded state reading standards at the same proportion as third grade Hispanic students throughout the state. For the total group of students, those in the district were less likely than those in the state to meet proficiency standards. Over time, the situation changed. By 2006 after three years of implementation (and when the third graders would have been exposed to RM since first grade), 95% of all students (both the total group and the Hispanic students) met or exceeded standards. The effect sizes associated with these changes, and controlling for simultaneous changes within the state, were .65 for the total group of students and.81 for the Hispanic students. Both of these effects were statistically significant.

12)SRA/McGraw-Hill. (2005j). *Phoenix inner-city students strive toward national reading average*. Columbus, OH: The McGraw-Hill Companies. (Note this was also listed by the WWC as SRA/McGraw Hill (n.d.m)

This case study reports data from Wilson Primary School, which serves pre-K through third grade students in Phoenix, Arizona. Ninety percent of the students were Hispanic, and 95 percent qualified for free and reduced lunch. The school began using Direct Instruction

programs in the fall of 1998. Data were presented in the SRA publication for Stanford Achievement Test/9 scores for third grade students from the spring of 1998, before implementation, through the spring of 2001. The data indicate that third graders' scores increased markedly with the new curriculum. While the average student scored at the 17th percentile before implementation, the average student scored at or just below the national average in later years. A cohort normative control group design was used to examine these data, comparing scores to national norms for the cohort tested before the program was implemented to scores after implementation. The percentiles were converted to Normal Curve Equivalent (NCE) scores for the calculations. The effect size associated with the change from 1998 to 2001 was .85, and the t-test assessing this change was statistically significant.

13) SRA/McGraw-Hill. (2005l). *Reading Mastery Plus helps Colorado school achieve AYP for first time*. Columbus, OH: The McGraw-Hill Companies.

This report presents data for Ivywild Elementary School in Colorado Springs. The school served 135 students in grades K through 5. Over 90 percent of the students received free or reduced lunch. Half were Hispanic and most of the other students were Caucasian. Forty-four percent of the students were English Language learners. The school began using the Direct Instruction program *Reading Mastery* in 2003-04. Data are given on the percentage of third grade students in the school and the state who scored at the proficient and advanced level of the Colorado Student Assessment Tests from the spring of 2003, before implementation, to the spring of 2005. Results indicate that students in Ivywild were more likely to score at the proficient or advanced level, even with controls for changes within the state as a whole. Patterns of change differed slightly for the three grade levels. The associated effect size was .63, which was statistically significant.[38]

14) SRA/McGraw-Hill. (2006a). *Cleveland school keeps Reading Mastery as curriculum core*. Columbus, OH: The McGraw-Hill Companies.

Louisa May Alcott Elementary in Cleveland, Ohio served 208 students in grades K-6. One hundred percent of the students qualified for free or reduced lunch. Slightly more than a third (35%) were African American, close to half (47%) were Caucasian, and most of the rest were Hispanic (12%). The school fully adopted the DI programs *Reading Mastery* and *Language for Learning* in 1998-1999. The study reports the percentage of fourth graders passing the state reading tests from the spring of 1998, when none would have been exposed to *RM*, through 2006. Beginning in 2003 fourth graders continuing in the school would have been exposed to *RM* and *LL* since kindergarten, thus having the maximum exposure. The data indicate dramatic increases in the percentage of fourth graders meeting state standards, from less than a third before implementation to well over three-quarters in

[38] Data were also given for grades 4 and 5, but these are not included in our analysis because they are not included in the WWC protocol for beginning reading.

2003 and later. There was a noticeable drop in 2006, which the school officials attributed to an influx of new students. Using the language of experimental design, this would suggest that the 2006 cohort was not strictly comparable to earlier cohorts. Thus comparisons of 1998 and 2006 as well as 1998 and 2005 to percentages for the state as a whole are included (a cohort control group with historical comparisons). Both comparisons indicate substantial effect sizes and statistically significant results, although results are smaller for the former comparison. The effect size for the 1998-2006 comparison was .63 and for the 1998-2005 comparison .94.

15)SRA/McGraw-Hill. (2006b). *DIBELS scores advance to grade level with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

Edgewood Academy in Fort Myers, Florida, is a Pre-K to fifth grade school with 700 students. Almost ninety percent (88%) of the students received free or reduced lunch, 34 percent were African American, 20 percent were Caucasian, and 40 percent were Hispanic. Almost 30 percent (29%) were ELL students. The school began using *Reading Mastery* in grades K-5 in August, 2006. Before that time they had found that the percentage of students who were assessed, using DIBELS, as being at grade level (or low risk) declined from the fall to mid-year testing. After implementing *RM*, however, this pattern altered. From fall to winter of 2004-05 (before implementation of *RM*), the percentage of students at "low risk" of failure (i.e. doing well) declined from 85% to 58%, a fall of 27 percentage points. After implementation, the percentage increased from 37% in fall to 43% in winter. One can assess the magnitude of these changes for the total group by a simple difference of difference in proportions test, essentially testing the null hypothesis that the difference between fall and winter in 04-05 equals the difference in 06-07 (see Blalock, 1979, pp. 234-236 for an example).[39] The corresponding effect sizes ranged from .04 to .72, with an average of .40. Results for grades one and two reached traditional levels of statistical significance.

16)SRA/McGraw-Hill. (2006e). *Native American school uses Reading First grant to implement Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

The Nay Ah Shing School in Onamia, Minnesota served 216 students in grades K-12. All of the students were Native American. At the start of the 2004-05 school year they began their implementation of the Direct Instruction program *Reading Mastery* in grades K-3. DIBELS data were provided for students in these grades at the start of the 2004 year (as implementation was beginning) and the end of the 2005-06 year. Thus, a cohort comparison design can be used to compare the percentage of students at benchmark at the beginning of the school year before significant exposure to *RM* to the percentage at benchmark at the end of two years of implementation. At that point both the kindergarten and first grade

[39] Results were given in the report for grades one to five, but only those for grades one to three are included in this analysis. Those for the upper grades also supported the efficacy of RM.

cohorts would have been exposed to *RM* throughout their schooling careers, while the second and third graders would have had 2 years of exposure (grades 1 and 2 or grades 2 and 3). The effect sizes ranged from .30 to 1.43, with an average value of .90. Effect sizes were larger and results reached standard levels of statistical significance only for the two lowest grades where students had received the most exposure to *RM*.

17)SRA/McGraw-Hill. (2007b). *Low-performing Kentucky school on its way to high-performing with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

Highland Elementary is a Pre-K to fifth grade school of 250 students in Hopkinsville, Kentucky. Ninety-two percent of the students received free or reduced lunch. Sixty-one percent were African American, and 38% were Caucasian. *Reading Mastery* was implemented in grades K-5 in October 2005. In 2005, before RM was implemented, the fourth graders had scores that were .84 of a standard deviation below those for students in the state as a whole, but by 2007, when they would have had RM for two years (third and fourth grade) their scores were .20 of a standard deviation above those for the state. The associated effect size was 1.04, indicating that, relative to changes in the state, scores in Highland increased by more than a standard deviation. This change was highly statistically significant. Note that, as with several other studies cited above, the comparison in this study involved an intervention group with dramatically lower scores at baseline, but post-test effect sizes that were dramatically greater than these baseline differences.

18)SRA/McGraw-Hill. (2007d). *Reading scores rise at Alabama elementary school with Reading Mastery Plus*. Columbus, OH: The McGraw-Hill Companies.

Elba Elementary School is a K-6 school in Alabama. Almost three-fourths of the students received free or reduced lunch. Slightly more than half of the students were African American, and most of the rest were Caucasian. The study reported the percentage of students in grade 3 who met or exceeded state reading standards on the Alabama Reading and Mathematics Test from the spring of 2005, before *RM* was implemented, to the spring of 2007. We also obtained the percentage of students in the state who reached these levels and used the cohort control group with historical comparisons design to calculate comparisons. The effect size associated with the change over time, relative to that in the state, was .11.

19)SRA/McGraw-Hill. (2007f). *Title I schools in North Carolina district meet all-state reading targets with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

This SRA report included information on two schools from Brunswick County, North Carolina that began using *Reading Mastery* in the 2004-2005 school year. Information was provided on the percentage of students scoring at or about the proficient level in reading with the End-of-Grade tests in grade from spring of 2004 through spring of 2007 for two schools. We obtained data on the corresponding percentages for the state. In 2004 Brunswick County students had no exposure to RM, but by 2007 they would have had three years of exposure

(beginning in grade 1). Over time, students in both schools were more likely to score at the proficient or higher level in reading. The effect sizes of these changes, relative to those in the state as a whole were .53 for one school and .36 for the other (average = .445). Both were statistically significant.

20)SRA/McGraw-Hill. (n.d.a) *Seattle school boosts reading scores with Reading Mastery curriculum*. Columbus, OH: The McGraw-Hill Companies.

This document reports data on a charter school where 94 percent of the students are African American and 89% are eligible for free or reduced lunch. The school has used *Reading Mastery* since its inception in 2000-2001. Since that time the percentage of fourth grade students in the school who exceeded state standards on the Washington State assessment grew from 47% to 85% in 2005-06. The increase in the state was from 66.1 to 81.2%. A cohort control group with historical comparison design was used to analyze the change and resulted in an effect size of .61, which was statistically significant. (Even though the data are on fourth graders, the WWC allows tests on a grade outside the range if the earlier instruction was in the studied program (WWC, 2012, p. 3). However, the WWC may object to including this comparison because the students in the first cohort had RM. They differed from those in the last cohort by the amount of exposure they had to the program. Those in the first cohort did not have *RM* in grades K-3, while those in the second cohort did. The conclusions of the impact of RM on achievement and the lack of impact of WWC standards remain when this study is omitted from the summary statistical analysis in the companion report.

21)SRA/McGraw-Hill. (n.d.b). *Anchorage school's diverse population flourishes with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

This document reports data on a public elementary school (K-6) with an ethnically diverse population and 88 percent free or reduced lunch rate. The school began using *Reading Mastery* in the 2001-02 school year. In the prior year (2000-2001), 27 percent of the third graders met or exceeded state reading standards, but by the end of 2005, 65 percent of the students were at that level. Using a cohort control group design, the effect size associated with the change was calculated as .83. The change was statistically significant.

22)SRA/McGraw-Hill. (n.d.n). "Nebraska District Outscores Peers Statewide, " pp. 14-15 in *Results with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

This report examines changes in the percentage of students meeting DIBELS benchmarks in grades K to 3 from before the implementation of RM (Spring 2004) to three years later (Spring 2007) in one mid-western district. Data are given for students in grades K to 3. A simple cohort control group design was used to analyze the data.[40] The effect sizes

[40] Because the DIBELS benchmarks do not change over time this design is equivalent to a normed comparison design. Note that other data from this site were used in other reports. In the multivariate analysis described in the companion report controls were included for multiple reports from a site.

associated with the changes ranged from .24 to .72, with an average of .50. All results were statistically significant.

23) SRA/McGraw-Hill. (n.d.p). *Success begins early at Alaskan elementary school*. Columbus, OH: The McGraw-Hill Companies.

This report provides data from one Alaskan school on spring DIBELS scores of kindergarten students before implementation of RM and after implementation as well as changes in third graders' scores on the state reading assessment. The school has a free and reduced lunch rate of 66 percent. Seventy percent of the students were Caucasian and 25 percent were Alaskan Natives. The effect sizes associated with the changes ranged from .13 to .87, with an average of .48.

24) Stockard, J. (2010). *The impact of Reading Mastery in kindergarten on reading achievement through the primary grades: A cohort control group design*. Eugene, OR: National Institute for Direct Instruction. (rejected by WWC for unacceptable design)

This short report, submitted by the author to the WWC in December 2010, used a cohort control group design. The short report included material explaining why a cohort control group design is considered by the standard methodological tradition to be especially appropriate for school settings. It reported data on 790 students, 652 who had RM in kindergarten and 138 who did not. There were no differences in literacy skills at the start of kindergarten ($d = .06$ on the DIBELS measure of Letter Naming Fluency and $d = -.03$ on Initial Sound Fluency). Yet, there were statistically significant differences in the DIBELS literacy measures at mid kindergarten (Nonsense Word Fluency, $t = 6.83$, $p < .0001$, $d = .58$) and at the end of the kindergarten year ($t = 3.83$, $p = .0002$, $d = .32$) . The results were included in Stockard, 2011 and Stockard and Engelmann (2010). The mixed model, meta-analysis results given in the companion model involve an analysis with design as the level 2 measure and an analysis with site as the level 2 (grouping) measure. This analysis (Stockard 2010) is treated separately from the other two in the former, but grouped with them in the latter. (See Technical Report 2014-3 for details.)

25) Watkins, T. B. (2008). *A comparative analysis of the effectiveness of Direct Instruction reading on African American, Caucasian, and Hispanic students*. Dissertation Abstracts International, 69(03A), 104-923.

This dissertation examined gains over time in vocabulary, comprehension, and total reading scores on the Iowa Test of Basic Skills for third, fourth, and fifth grade students enrolled in an Arkansas elementary school. Data were reported on fall and spring scores of the total group of students, as well as separately for those in different race-ethnic groups (Caucasian, African-American, and Hispanic). While the dissertation focused on only the pretest-posttest changes, it would be possible to analyze these data as a norm comparison design. At the time of this writing we did not have the necessary data, but will conduct the analysis when

the data are available. It was probably rejected by the WWC because it lacked an intact control group.

# Appendix D
# Studies that Failed to Meet WWC Standards of Evidence at Step Three

This appendix describes studies of *Reading Mastery* that were rejected at step three of the WWC's review. All of these studies employed a pretest-posttest control group design, but were rejected for failing to meet the WWC's standards of evidence. The discussion parallels that of the text by examining each set of standards separately. Within the discussion of each standard the studies where the WWC decision was determined to be appropriate are described first followed by a description of the studies where that decision could be questioned if standard methodological criteria were used.

## Criteria of Group Equivalence

The WWC rejected 13 studies because the intervention and control groups did not meet the standard for equivalence at pretest. Our analysis found that seven of these decisions were appropriate, albeit not always because of the reasons used by the WWC. The decision regarding six of the articles could be legitimately questioned.

### Appropriate decisions to exclude

1) Kamps, D., Abbott, M., Greenwood, C., Wills, H., Veerkamp, M., & Kaufman, J. (2008). Effects of small-group reading instruction and curriculum differences for students most at risk in kindergarten. *Journal of Learning Disabilities, 41*(2), 101–114.

This study combined results of *Reading Mastery* with other programs, so it is impossible to separate out the impact of *RM*. There is another article by Kamps and associates that does examine *RM* as a separate curriculum (Kamps, et al., 2003, listed in Appendix B), but it was not included in the WWC's listing. Kamps, et al. (2008) will be omitted from our forthcoming meta-analysis, Kamps, et al., (2003) will be included.

2) League, M. B. (2001). *The effect of the intensity of phonological awareness instruction on the acquisition of literacy skills*. Dissertation Abstracts International, 62(10), 3299A.

This study compared three groups of students. One had *Reading Mastery* plus phonological training, one had *Reading Mastery* alone, and the third had some phonological training by computer. However, the groups differed in the amount of intervention time that they had. The WWC's decision to omit the study appears appropriate and it will be omitted from our meta-analysis.

3) McCollum, S., McNeese, M.N., Styron, R., & Lee, D.E. (2007). A school district comparison of reading achievement based on three reading programs. *The Journal of At-Risk Issues, 13*(1), 1-6; The article is based on McCollum-Rogers, S. (2004).

*Comparing Direct Instruction and Success for All with a basal reading program in relation to student achievement*. Dissertation Abstracts International, 65(10), 3642A (UMI No. 3149920) .

This work compared the average Wide-Range Achievement scores of third grade students in three Title I schools in a Caribbean territory of the United States. One school used Success for All (SFA), one used DI programs including *Reading Mastery*, and one used a basal reader called *Literature Work*s. Training and regular coaching and support through both site visits and telephone meetings were provided to the SFA school. No coaching or instructional supports were provided to the other two schools. No pretest data were provided, nor were data provided regarding the demographic characteristics of the three schools. Results reported showed that the students in the SFA and basal reader school had higher average scores than those in the DI school (effect sizes of -.37 in comparison to SFA and -.18 in comparison to the basal reader). Given the lack of any information regarding equivalence of the schools at baseline, as well as the differences in support given to teachers, the WWC's decision to exclude the study appears to be reasonable. Note that, given the lack of pretest data, it is unclear why this study was not rejected at step 2.

4) Neely, M. (1995). The multiple effects of whole language, precision teaching and Direct Instruction on first-grade story-reading. *Effective School Practices, 14*(4), 33–42.

This study compared students in three successive third grade classes (cohorts): those who received 1) a whole language reading program, 2) whole language plus precision teaching, and 3) whole language plus precision teaching plus *Reading Mastery*. As would be expected from other literature, the students in the third group (who received *Reading Mastery*) had the highest reading skills and gains during the year. The author noted that the year three program "was three-and-a-half times more effective" than the other years, thus indicating a large "value added" effect of the program. However, because there was no group of students that received only *Reading Mastery,* it was not possible to calculate effects of the sole contribution of *RM*, and it was appropriate for the WWC to exclude it from their analysis.

5) Ryder, R. J., Sekulski, J. L., & Silberg, A. (2003). *Results of Direct Instruction reading program evaluation longitudinal results: First through third grade 2000–2003*. Retrieved from: http://www.uwm.edu.

This study had two different parts. One part compared students who had either *RM* alone or *RM* combined with other programs with students with another program. Thus there was no way to separate the effects of RM from that of the combination of treatments. In the other part of the study the progress of low ability first graders who received a combination of RM and other programs was compared with the progress of high ability first graders who received only the other programs. Again there was no group that only received RM. In addition the students differed at baseline in ability. The decision to omit this study was appropriate.

6) Thames, D., Kazelskis, R., & Kazelskis, C. R. (2006, November). *Reading performance of elementary students: Results of a five-year longitudinal study of direct reading instruction*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Birmingham, AL.

This unpublished work examined differences in reading achievement of students who began instruction in *Reading Mastery* at kindergarten with those who began instruction in later grades and found few significant differences. Data came from 6 elementary schools in one district. All of the groups in the analysis were eventually exposed to *RM*. None of the data included in the study allowed a comparison between students' who had only had *RM* with those who did not. Instead the analysis focused on comparisons of students who began the program at kindergarten and those that started at later points, with no inclusion of comparative data at the end of kindergarten. Thus, the decision to exclude the study appears reasonable. This paper will not be included in the meta-analysis.

7) Thomson, B., & Miller, L. D. (1991). Pilot study of the effectiveness of a Direct Instruction model as a supplement to a literature-based delivery model; Traditional teaching to a whole language: A focus on instructional routines. *Florida Educational Research Council Research Bulletin, 23*(2).

This paper examined the use of *RM* as a supplement to another program for first graders . Results indicated that the group given *RM* had significantly lower scores on pretest measures, but had scores that were equal to the comparison group on posttest. Because there was no group that had *RM* by itself, the decision to exclude the study appears appropriate.

## Potentially Inappropriate Decisions to Exclude because of Issues with Group Equivalence

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

Brent and associates compared scores on the Comprehensive Test of Basic Skills (CTBS) of students who had *Reading Mastery* in first and second grade with those who had a "traditional" basal program. Four classrooms were included in the analysis, two with DI and two with the traditional program. Data were reported for the fall of second grade and the spring of second grade. No data were reported for first grade because the district did not administer standardized tests at earlier years. Thus, there were no pretest data for the start of first grade. The comparisons of data at the start of second grade involved unadjusted means, while the comparisons of data from the end of second grade adjusted for scores at the start of that year. All effect sizes favored the *Reading Mastery* groups. Effect sizes at the start of second grade (after one year of *RM*) ranged from .03 to .20, with an average of .12. Effect sizes at the end of second grade, after two years of exposure and with scores adjusted to reflect differences at the start of the year, ranged from .85 to 1.09, with an

average of .97. To parallel the requirements of the WWC for pretest scores, only the results for the end of Grade 2 are used in our analysis, for they involve pretest scores and control for the differences in these scores in the analysis. Another paper by two of the authors (Brent and DiObilda, 1993, listed in Appendix B and included in a list of references sent by NIFDI to the WWC) could have been appropriate to consider.

27) Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology, 47*, 187-214.

This quasi-experimental study was supported by grants from IES, the body that supports the WWC, and the National Institute for Child Health and Human Development (NICHHD). The authors examined growth in reading achievement during one school year of over 30,000 students in grades one to three who were randomly selected from almost 3000 classrooms. They compared changes in oral reading fluency from fall to spring of students in *RM* and five other curricula using growth curves and hierarchical linear modeling. The authors reported descriptive statistics at baseline on oral reading fluency for each of the groups in the analysis (p. 192) and for the total group. Of the 15 possible comparisons of a curriculum with *RM*, three exceeded the .25 criterion set by the WWC. On average, the *RM* sample differed from the other groups by .12 of the total s.d., while the absolute value of the deviations ranged from .03 to .40 of the total s.d.. The fact that three of the fifteen comparisons exceeded the established level of .25 s.d. apparently resulted in the study being rejected. Interestingly, the differences that were larger than the criterion appeared in only two of the three grade levels that were analyzed, and the authors reported results separately for each grade. It is unclear why the WWC chose to ignore all of the results rather than accepting those for the one grade level where the differences fell within the designated level. In addition, as explained in detail in the companion Technical Report (2014-3), the probability that at least one of the comparisons would involve differences than .25 s.d. would, by chance, be relatively high.

The authors summarized their findings as follows:

> Overall, students in the *Reading Mastery* curriculum demonstrated generally greater overall ORF growth than students in other curricula. Also, they more frequently met or exceeded benchmarks for adequate achievement in first, second, and third grade (p. 209).

Effect sizes were calculated comparing *RM* with each of the other 5 curricula for each of the 3 grades for the total group and separately for high and low SES students, resulting in 15 effect sizes for each grade and 45 effect sizes in total.[41] The calculated effect sizes ranged

[41] The intercept in the growth models (from Table 3, p. 198) was used as the estimate of the posttest mean. The estimated common standard deviation for each comparison was calculated from data obtained from Table 2 (p. 192).

from 0 to .40, with an average of .23. Similar results occurred with both high and low SES students, but those were omitted from this analysis for the sake of simplicity.

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

These data are also reported in Rightmyer, E. C., McIntyre, E., & Petrosko, J. P. (2006). Instruction, development, and achievement of struggling primary grade readers. *Reading Research and Instruction, 45*, 209–241. This article was not included in the WWC listing of works reviewed.

McIntyre, et al. (2008) and Rightmyer, et al. (2006) report on the same data set. First grade teachers in 12 to 17 different schools and 37 classrooms were asked to nominate "struggling" students for inclusion in the study. There were two to five students in each class. Gains in reading achievement over time were compared for students in schools using *Reading Mastery* (n = 56 in Grade 1) and those using other models (total n = 52 in Grade 1), all described as "non-scripted": *Breakthrough to Literacy, Early Success, Four Blocks,* and *Together We Can.* Clay's Hearing Sounds in Words was used to measure gains from fall to spring of first grade and the Flynt Cooper Informal Reading Inventory was used to measure gains from the beginning of first to end of second and the beginning of second to the end of third. Pretesting occurred in September and posttesting in May. Gain scores, rather than pretest and posttest scores, are reported. The results reported by McIntyre, et al. (2008) and Rightmyer, et al. (2006) were combined. The 13 effect sizes calculated from their data ranged from –.76 to .77, with an average of .11. While the use of gain scores controls for differences at pretest, the lack of actual pretest scores violates the WWC criterion and that is no doubt why the study was omitted. Although not mentioned by the WWC, the study could also be omitted because it focused on "struggling" readers, rather than general education students.

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

Note that this study was also rejected at step 2 of the WWC's evaluation. Two different types of analysis were used in the article, and that may be the reason that the study was included at two different stages. To parallel the WWC report, we discussed the article in both Appendix C and D. In the meta-analysis results, given in the companion report, results reported in Appendix C and D are kept separate when design is used as the Level 2 variable and combined when site of the study is used as the Level 2 variable.

O'Brien and Ware reported results of an evaluation of the implementation of *Reading Mastery* and *Open Court* in the Fort Worth Independent School District, replacing a whole language program that was used in earlier years. Implementations began with the lowest

performing schools in the district and principals chose the reading program that they would implement. In the first year, 32 schools participated; 18 principals chose to use *RM* and 14 chose to use *OC*. In the second year, 2 additional schools chose *RM* and the remaining 25 used a slightly different version of *OC*. Some schools continued to use their traditional whole language approach. The students in the traditional program were less likely to be minorities or to qualify for free or reduced lunch. The students in the *OC* and *RM* programs had similar socio-demographic characteristics. Training and support was provided for both the *OC* and *RM* implementations. Implementation fidelity was monitored.

Results of two assessments were reported. Those with the first assessment were discussed in Appendix C with the analysis of studies rejected at step 2. The second assessment was the Stanford Achievement Test (9th edition). Because the groups of schools differed in socio-demographic characteristics, value added regression analyses were used to control for prior test scores and demographic variables. (The article did not report the standard deviation for each group in the analysis, and thus we could not test the actual extent to which these fall scores varied.) To control for prior scores and demographic characteristics spring NCE scores on the SAT were regressed on fall scores and demographics. Data reported here are for the kindergarten students (Table 8, p. 190). The resulting regression equations were used to calculate effect sizes comparing *RM* students to those in the two *Open Court* curricula and in the traditional whole language approach for the total group and for sub-groups of African American, Hispanic, and White students.[42] Effect sizes ranged from .03 to .60 with an average value of .26. Given the differences between the schools in demographic characteristics, it is quite likely that the groups would have been unequal at pretest, and thus violated the WWC criteria. Unfortunately, the authors did not include sufficient pretest information to determine the extent of these differences, although they were, of course, controlled in the multivariate analysis.

30) Stockard, J. (2011). Increasing reading skills in rural areas: An analysis of three school districts. *Journal of Research in Rural Education, 26*(8); and
Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's Reading Mastery. *Journal of Behavior Assessment & Intervention in Children, 1*(1), 2–24. Study B

The second article (Stockard and Engelmann, 2010) included results from two separate studies, one of which involves a subset of the data used in the Stockard (2011) analysis. Because both used the same design, the results were combined in the meta-analysis that used design as the level two measure. Both reports used a cohort control group design to compare growth of reading skills of students who had *RM* from the beginning of kindergarten (full exposure cohorts) with other students. For the 2011 piece, data on almost

[42] Because NCE scores have, by definition, a mean of 100 and a standard deviation of 21.06, the unstandardized regression coefficients were divided by this standard deviation to obtain the effect size.

1700 students were obtained from three districts, including five separate schools. The author provides an extensive discussion (p. 5) on how the design counters the "reactive effects" that are common when employing random assignment in institutional settings and presents data showing that the students in the two cohorts were equivalent in three measures commonly used to assess risk of learning difficulties: receipt of free or reduced lunch and minority status. Descriptive data were given on DIBELS nonsense word fluency (NWF) from the beginning of kindergarten to fall of second grade and on oral reading fluency (ORF) from winter of first grade through spring of third grade. Linear growth models were used to examine changes in scores over time, controlling for students' at-risk status. Results indicated that, by the middle of kindergarten, those in the full exposure cohorts had significantly higher skills than students in the other cohorts, who did not have RM in kindergarten, and scores that were equal to or higher than a national sample. The effect sizes associated with comparisons with the cohort with no exposure to RM were .76 in winter and .73 in spring. Those associated with comparisons to national data were .65 and .69 (page 8). Note that because these comparisons are limited to kindergarten students they only include students who either did or did not have exposure to *RM*.

Similar data were given in the 2010 piece, for fewer schools. Data were provided on pre-literacy skills using assessments from the DIBELS measures: initial sound fluency and letter naming fluency. There were no significant differences between the cohorts, and the differences between the means ranged from .07 to .08 of the common standard deviation. This value is slightly above the WWC's cut-off criterion for analysis without statistical controls. But, as in the 2011 analysis, growth curve models were used to analyze the data, controlling for students' initial skills. The results parallel those discussed above for the larger sample in the 2011 article.

One of the articles did not include pretest data in the form of start of kindergarten scores, and the WWC may well have concluded that the use of demographic risk factors was unsuitable as a pretest control. Another report of these data was rejected for inclusion at step 2 (Stockard, 2010, number 24 in Appendix C) because of the nature of the design. That report does include the pretest scores, showing as noted immediately above that they were well within the WWC limits, albeit with statistical controls required. Interestingly, the author of the study was not contacted to see if such information were available, even though the WWC protocols indicate that such information will be solicited when it would aid a review. (The results of study number 24 in Appendix C are combined with those of these articles in the mixed model analysis in the companion report that uses site as the unit of analysis.)

## Design Included Unacceptable Confound

Ten studies were rejected at step three for having an unacceptable confound. The present analysis determined that six of these decisions were appropriate, but found reason to question four of the decisions. Studies in both groups are described below.

## Appropriate to Omit

Algozzine, B., Wang, C., White, R., Cooke, N., Marr, M. B., Algozzine, K., Helf, S. S., & Duran, G. Z. (2012). Effects of multi-tier academic and behavior instruction on difficult-to-teach students. *Exceptional Children, 79*(1), 45-64.

In this study *RM* was used for all Tier 3 students and for some Tier 2 students, but results were not reported separately for these students. Thus, the decision to omit the article is appropriate for data were not reported for a group that did not have *RM*.

Foorman, B. R., Chen, D. T., Carlson, C., Moats, L., Francis, D. J., & Fletcher, J. M. (2003). The necessity of the alphabetic principle to phonemic awareness instruction. *Reading and Writing, 16*(4), 289–324.

In this study the data regarding *RM* were combined with that of other curricula, so the decision to omit the study is appropriate.

Kamps, D., Abbott, M., Greenwood, C., Arreaga-Mayer, C., Wills, H., Longstaff, J.,...Walton, C. (2007). Use of evidence-based, small-group reading instruction for English language learners in elementary grades: Secondary-tier intervention. *Learning Disability Quarterly, 30*(3), 153–168.

In this study results with *RM* were combined with those of two other curricula, so the decision to omit the study is appropriate.

Kamps, D. M., & Greenwood, C. R. (2005). Formulating secondary-level reading interventions. Journal of Learning Disabilities, 38(6), 500–509.

This article combines the results of *RM* with other curricula, so the decision to omit the study is appropriate.

O'Connor, R. E., Harty, K. R., & Fulmer, D. (2005). Tiers of intervention in kindergarten through third grade. *Journal of Learning Disabilities, 38*(6), 532–538.

The intervention in this study used only selected portions from RM and combined these selections with other material. Again, the decision to omit the study is appropriate.

Trout, A. L., Epstein, M. H., Mickelson, W. T., Nelson, J. R., & Lewis, L. M. (2003). Effects of a reading intervention for kindergarten students at risk for emotional disturbance and reading deficits. *Behavioral Disorders, 28*(3), 313–326.

This study combined *RM* with another reading program. Thus, the decision to omit the work is appropriate.

## Potentially Inappropriate Decisions to Exclude

31) Carlson, C. D., & Francis, D. J. (2002). Increasing the reading achievement of at-risk children through Direct Instruction: Evaluation of the Rodeo Institute for Teacher

Excellence (RITE). *Journal of Education for Students Placed At Risk, 7*(2), 141-166. Reprinted in *Journal of Direct Instruction, 3*(1), 29–50.

This study examined the impact of implementation of *Reading Mastery* in the Houston public schools, comparing changes in 20 schools that implemented the program with those in 20 comparison schools. The schools were matched on demographic characteristics (receipt of free or reduced lunch, ethnicity, and English language proficiency) and average achievement (meeting the state mandated reading performance requirements) (p. 144). The matching occurred in the first year of the program. Within the pool of possible matches the school that was geographically closest to the RM school was chosen. Because the program was gradually implemented over a few years the authors were able to examine the longitudinal impact of the program and the extent to which greater exposure to the program increased scores. Assessments varied over the years, with individual assessments (the Woodcock Reading Mastery Test) used only in the first year (for kindergarten students in fall and spring) and district-mandated assessments in later years: the teacher administered Texas Primary Reading Inventory (TPRI) in winter and spring (kindergarten), the Stanford Achievement Test Word Reading and Reading Comprehension scales (grades 1 and 2, word reading only for kindergarten), and the Texas Learning Index from the state assessment program (grade 3), for one year. Teachers were trained in the use of the program, given in-class support, and observed to determine fidelity to the program. Given the four year duration of the study and the multiple grades and schools involved, the probability of issues with attrition and comparability of samples was not small. The authors report (pp. 149-50) special efforts to ensure comparability and test for the impact of differential attrition.

The authors used multilevel modeling techniques in all analyses, with individual student as the first level and classroom (or teacher) as the second level. This adjusts for the lack of independence among children studying within the same classroom. Analyses within each grade also adjusted for the number of years students were exposed to the program (p. 150). Pretest data were available for only one year (1997-98) and only for kindergarten students. The authors reported, "Results indicate that groups did not differ in the fall, $F(1,412)=.08$, $p<=.78$; and there was no evidence for heterogeneity of regression."[43] Seventeen effect sizes could be computed, and they ranged from .07 to 7.85 in value, with an average of .79. Eleven of the values were greater than .25. The extraordinarily large effect size occurred for the WJRM Test of word identification for kindergarten students. When that value is omitted, the range is .07 to .66, and the average is .35, still well above the cut-off for educationally significant results.

The study was also rejected for inclusion in the 2008 WWC report on *Reading Mastery* and Beginning Reading because of a supposed confound. Given the quality of the study, the

[43] Because there are only two groups involved, this F value can be converted to a t-value (t = the square root of F, df = $df_2$ of the F. The resulting t value of .2828 is equivalent to an effect size of .03 (the difference of group means in s.d. terms), and thus is well within the requirements of the WWC for group equivalence.

reason for its exclusion was of special interest to the NIFDI Research Office. The only intervention in the experimental schools was *Reading Mastery*, and as with the other studies, the results strongly favored this curriculum. After scouring the article, we found one possible explanation for the WWC's determination that there was a confounding effect. 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).[44] A September 8, 2008, communication from Mathematica confirmed 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).

The NIFDI Research Office responded to this judgment as follows:

> In reality, strong classroom management is an integral element 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. 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 (Stockard, 2008, p. 11).

There are other reasons that the WWC could decide to omit the study. As noted above, pretest data were available for only one grade and cohort, although the differences were well within the limits required by the WWC and there were extensive statistical controls. However, the results were all obtained from one district, albeit a very large district and involving 40 schools, and thus the study could be rejected with the "one unit" criterion.

32) Jones, C. D. (2002). *Effects of Direct Instruction programs on the phonemic awareness abilities of kindergarten students*. Dissertation Abstracts International, 63(03), 902A.

[44] 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.

In this study 36 kindergarten students in four classrooms in one school were randomly assigned to treatment groups. All of the students were described as being "at risk" and having scores in the bottom quartile of the Brigance Diagnostic Inventory of Basic Skills. However, students who were developmentally delayed or received special education services were excluded. (Forty students were originally in the study, but attrition dropped the n to 36.) Sixteen of the students were in the school's usual instructional program, with seven receiving their traditional classroom reading instruction and nine receiving one-on-one tutoring in a phonological awareness program developed and taught by the school staff. The other 20 students were assigned to Direct Instruction training. Half of these students received group instruction in *Reading Mastery*, and half received one-one-one tutoring guided by the book *Teach Your Child to Read in 100 Easy Lessons*. This book is written by the author of RM and incorporates all of the strategies used in RM but in a format that is appropriate for tutoring. Students in the intervention group received tutoring from graduate students from a nearby university and those in the control group received tutoring from school staff. Instruction occurred daily for one hour for four weeks. Phonological skills were assessed with the Comprehensive Test of Phonological Processing. Analysis of covariance was used to control for pretest scores when assessing the differences between the groups. Results indicated a statistically significant difference between the groups and an associated effect size of .49 (p. 94). Because there was differential attrition between the two groups, the author assessed the possibility that such differences might have arisen from the treatments and found no evidence for that possibility.

The WWC provided no discussion of why they rejected the study. One possible explanation could involve the combination of *RM* and *Teach Your Child to Read (TYCR)*. This combination parallels, of course, the combination of tutoring and whole class instruction in the control group. The curricular content of *RM* and *TYCR* is very similar. In addition, the author included a test of homogeneity of slopes to test if there was interaction of the treatment and the covariate (the pretest scores), and the results were insiginificant ($F=0.03$, $p = .87$). It is also possible that the study was rejected because the DI groups were taught by graduate students and the control groups were taught by school staff. Finally, it is possible that the study was rejected because limited data on pretest scores were given, even though subjects were randomly assigned and the analysis of covariance clearly controlled for any differences that might have occurred. Finally, it is possible that the WWC would reject the study because data came from one school, even though multiple teachers were involved. It is not clear if the WWC requested additional information from the author, such as a breakdown of results by groups.

33) Mac Iver, M. A., & Kemper, E. (2002). The impact of Direct Instruction on elementary students' reading achievement in an urban school district. *Journal of Education for Students Placed at Risk, 7*(2), 197–220.

The article reports on a comparison of 6 schools in the Baltimore City Public School System (BCPSS) that implemented *Reading Mastery* as part of a whole school intervention project with 6 other schools matched on demographic characteristics.[45] Outcome measures were scores on the reading comprehension subtest of the Comprehensive Test of Basic Skills (CTBS) and a curriculum-based measure of oral reading fluency. Covariates used as controls included the Peabody Picture Vocabulary test (for those entering the study in Kindergarten) and CTBS scores from previous years. The authors report effect sizes, controlling for pretest scores and demographic variables.[46] Results are given for two cohorts: one where the intervention group began *RM* in kindergarten and another in which the intervention group began *RM* in grade 2.

This study was rejected at step 3 because "the intervention was not implemented as designed." This determination is understandable, given a lengthy discussion (pp. 200-202) of problems in implementation fidelity at the schools and variations across sites. In this discussion the authors justify their decision to retain data on schools that were judged by the developer of *RM* to have used the program with lower levels of fidelity, They justify their inclusion of schools that did not implement as the developer deemed necessary, arguing that "a process evaluation cannot disregard the cases that fail to meet this standard or the reasons they fail to meet it" (pp. 202).[47] However, to address some of the concerns

---

[45] Note that this is the same population used by Butler (2003) and discussed in Appendix C. Another article (Stockard, 2011a, listed in Appendix B) also uses this data set, but it was not included in the WWC list. It examines a longer time period than the Mac Iver and Kemper article and also includes more extensive analysis of the impact of implementation fidelity on results. The Butler and Mac Iver and Kemper articles are combined in the mixed model analyses in the companion report when site is used as the level 2 variable.

[46] The average PPVT score (NCE) at the start of Kindergarten was 29.7 for the intervention group and 31.6 for the comparison group. This difference is .09 of the standard deviation of 21.06 for NCE scores, favoring the comparison group. The value was within the WWC limitations of acceptability if statistical controls are used, as the authors did.

[47] Interestingly, the authors also appear to assert that if a program (such as *RM*) requires that students learn to mastery one cannot have a valid measure of its effectiveness: "…

> because the program [RM] also requires that students learn to mastery before teachers are able to progress further with lessons, it appears that the developer's definition of successful implementation incorporates a student outcome component. This blurring of distinctions between implementation and student outcomes (even if not the same outcome measure as used in the evaluation) complicates the evaluation process, especially if the developer claims that implementation is low because a certain number of lessons were not mastered in kindergarten." (p. 201)

In other words, the authors appear to suggest that because successful implementation involves ensuring that students learn (master) the content and that they do so at a pace that enables children to catch up with grade level peers, appropriate implementation makes it difficult to have an independent assessment of learning outcomes. In the view of the author of this technical report, this logic seems to present a "damned if you do and damned if you don't" scenario. If one teaches to mastery and at a pace needed to catch up with others (a very central element of Direct Instruction), the logic of teaching to mastery as part of the program can challenge the "objective" nature of other assessments. On the other hand, if you don't follow the requirements of the program (e.g. do not teach to mastery and do not accelerate pace of learning) students will not learn as much and scores on outcome tests will be lower. A logic that seems to argue for low fidelity of implementation in order to avoid a "blurring" of implementation fidelity and outcome measurement is extraordinarily difficult to understand, for such logic would appear to result in a very invalid test of the intervention.

regarding fidelity, results were reported separately, at some points, for data from schools that worked directly with the developer of *RM* and may have had higher levels of fidelity and those that did not.

The authors report a large number of results, some of which control for demographic characteristics and some of which do not. For the current purposes we have used the effect sizes obtained from regression equations that controlled for demographic variables and pretest scores (Table 1, p. 203). The effect sizes with all schools included ranged from -.04 to +.21, with an average of 07. When only the schools with higher fidelity were included, the effects ranged from 0.0 to .21, with an average value of .11. Given that the WWC rejected the study because of apparent concerns with fidelity, we have only included the higher fidelity results in our analysis. This would, supposedly, be the results the WWC would have chosen to include if they had retained the study. The study appears to meet the WWC criteria for group equivalence at pretest (albeit with a need for statistical controls). However, the WWC might choose to reject the study because all the data came from one district, thus violating the "one unit" rule.

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

This study compared the reading performance of 31 students from a low income rural area in the southeast who were taught with a traditional basal approach (using the Houghton-Mifflin Reading Series) or with *Reading Mastery*. Students were nominated by their teachers as needing extra help and were randomly assigned to one of the groups. There were no significant differences between the groups before the intervention on either the Otis Lennon School Abilities Test or the total Reading Score of the Woodcock Johnson Reading Mastery Test. Students in the comparison classrooms were taught by their regular teachers (2 classrooms) and a university practicum teacher. Students in the *RM* group were taught by Masters degree practicum students, again with four teachers in total. Students were taught daily for the entire year. The HM teachers had been using their program for a number of years. The *RM* teachers were given training before they began. All teachers were observed "at least weekly" (p. 116) and given feedback to enhance their fidelity to the respective programs. Analysis of scores on the WJRM test at the end of the year indicated that students with *RM* had significantly higher scores on the total battery as well as subtests of passage comprehension and word identification. Effect sizes were very large, ranging from 1.11 to 4.01, with an average of 2.44.

We are unsure why this paper was judged to have a confound. Both programs were implemented as designed with checks, at least weekly, to ensure that there was program fidelity. Both groups had two teachers involved, although the comparison group had more experienced teachers. This, however, would work against the possibility that the *RM*

students would have higher scores. Training of the RM teachers was explicitly mentioned, but one would assume that the basal teachers, given their longer years of experience, would also have had such training. It is also probable, however, that the study would have been rejected for a lack of group equivalence. Differences between the groups at pretest (calculated from the t-values) were .14 of a standard deviation for the Otis-Lennon, with scores favoring the experimental group. Differences at pretest on the Woodcock Johnson were -.33 of a standard deviation, with scores favoring the control group. The average difference was -.10, favoring the control group. However, the WWC routinely rejects a study in which any pretest difference exceeds .25 of a standard deviation, even if the differences average out to below the criterion. The WWC could also reject the study because it occurred in one school (albeit the only school in a small rural, very low income area), and thus violating the one unit rule.

## One Unit Per Condition

Seven articles were rejected under the "one unit per condition" standard. The decision to exclude appears to be appropriate for three of these studies, but arguably inappropriate for four of them. All the articles are described below.

### Appropriate to Exclude

Ryder, R. J., Burton, J. L., & Silberg, A. (2006). Longitudinal study of Direct Instruction effects from first through third grades. *Journal of Educational Research, 99*(3), 180–191.

This study appears to use the same sample that was used for the Ryder, et al (2003) analysis discussed above. There was no group of students in the comparison that received only *RM* for reading instruction. Thus, the decision of the WWC to exclude the study appears to be appropriate, although perhaps not for the given standard.

Wiltz, N., & Wilson, G. P. (2006). An inquiry into children's reading in one urban school using SRA *Reading Mastery* (Direct Instruction). *Journal of Literacy Research, 37*(4), 493–528.

This article reports a descriptive case study of 27 second graders in one school, designed to qualitatively understand the way in which they used reading strategies and comprehended what they read. All students in this study had *RM*, and there was no comparison to another curriculum or to normative data. The study will not be used in our meta-analysis, and the decision of the WWC to exclude the study appears appropriate.

Wills, H., Kamps, D., Abbott, M., Bannister, H., & Kaufman, J. (2010). Classroom observations and effects of reading interventions for students at risk for emotional and behavioral disorders. *Behavioral Disorders, 35*(2), 103–119.

This study combined results of *Reading Mastery* with other programs, so it is impossible to separate out the impact of *RM*. We will omit the Wills, et al. (2010) piece from our meta-analysis. The decision by the WWC to omit it is appropriate.

## Arguably Inappropriate Decisions to Exclude

35) Ashworth, D. R. (1999). Effects of Direct Instruction and basal reading instruction programs on the reading achievement of second graders. *Reading Improvement, 35*(4), 150–156.

This study compared the reading achievement, measured by the Iowa Test of Basic Skills, of two classes of second grade students, both taught by the same teacher. In the first year (n=20 students) the teacher used the basal reader that the district had been using for a number of years. In the second year (n=16 students) the teacher used *RM*, with implementation and coaching support from an external consultant. The author reported no difference between the two groups in scores on the Georgia Kindergarten Assessment Program and concluded that they had equivalent intellectual ability at the start of their schooling. Posttest scores on the ITBS indicated significant differences between the groups. The effect size associated with differences on the total score of the ITBS was 1.60.[48]

The WWC did not give details on why this study was rejected, but it probably relates to the fact that the same teacher and the same school were involved. Yet, the design of this study, which could be termed a pretest-posttest control group design, is one that is recommended by the CCSS tradition for schools. Having the same teacher administer a treatment to different cohorts is a standard way to counter the confounding effect of one teacher per treatment. Although the author reported that there were no differences between the groups in their scores on the ability test given at the start of kindergarten, numbers for these scores were not reported and thus it was not possible to check the extent to which the groups actually differed. This could be another reason that the WWC would reject the article.

36) Green, A. K. (2010). *Comparing the efficacy of SRA Reading Mastery and guided reading on reading achievement in struggling readers*. Dissertation Abstracts International, 71(11A), 3969.

This dissertation examined growth in reading achievement of second grade students enrolled in two different schools with very similar demographic characteristics in a rural South Carolina community. The students were all characterized as "struggling readers," scoring at least one year behind grade level on the NWEA Measures of Academic Progress at the beginning of the school year. Students in one school (42 students, 5 classrooms) had Guided Reading as the curriculum, and students at the other school (24 students, 3 classrooms) had *Reading Mastery*. The Measures of Academic Progress were assessed at fall (pre-test), mid-year, and in the spring. Instruction occurred for 7 months. Results indicated that the students in *RM* had less progress over time than students in the Guided Reading Program. The associated effect sizes were -.51 for the mid-year testing and -.55 for

[48] This effect size was calculated from the t-ratio reported in the article. Mean scores were given for the three subtests (vocabulary, comprehension and language), all of which showed the same pattern. Standard deviations were not, however, reported for these values.

the spring testing. The fall pretest scores also showed a slight advantage for the Guided Reading group (d=-.12). Analysis of covariance was used to control for these pretest differences.

It is possible that the WWC chose to omit this study using the "one unit" standard because the students came from two schools. However, multiple teachers were involved in both the intervention and comparison condition, pretest differences fell within the WWC established range, and multivariate statistics were used to control for initial differences between the groups. It will be included in our meta-analysis.

37a and b) SRA/McGraw-Hill. (2009). *A report on the effects of SRA/McGraw-Hill's Reading Mastery, Signature Edition: A response to intervention solution*. DeSoto, TX: Author.

This report describes the results of implementing *RM* with students in grades 2 to 4 at an inner city school in New York City. The school had a very high poverty rate and a large minority student population. The authors reported that the principal and the researchers initially intended to use a randomized control group design. However, "as word of an impending reading intervention study spread throughout the school, an overwhelming majority of the teachers asked the principal if they could participate in the study" (p. 3). The principal asked that the study be modified to include all of the interested teachers.[49]

Two types of data were gathered: the percentage of students scoring at the proficient level or higher on the New York ELA assessment and the percentage of students at grade level on the Rigby Reading Diagnostic and Evaluation System, a standardized assessment adapted from the Metropolitan Achievement Test. Two different designs were used. A cohort control group design was used to analyze both the EAL and the Rigby data. The percentage of third grade students who met proficiency standards in 2008 before the implementation began (29.9%) was compared with the percentage who met the standards in 2009 (87.5%). In addition the percentage of students in grades two through four that were at grade level in the spring of 2008 was compared with the percentage at grade level in the spring of 2009. In all four comparisons the percentages were strikingly higher for the 2009 data. Effect sizes ranged from .73 to 1.46, with an average of 1.13. A second analysis used a pretest-posttest norm comparison panel design. This analysis compared the percentage of third graders in 2008 who scored at the proficient level on the EAL (29.9%) with the percentage of students in this cohort who scored at the proficient level in 2009 when they were fourth graders

[49] Data were given in graphs from four peer schools, selected on the basis of similar demographics and past academic performance. All of the peer schools had higher scores on the New York English Language Arts (ELA) exam than the target school, and these differences were larger than the WWC cutoff, ranging from .24 of a s.d. to 1.15 s.d. Results shown in a graph indicate that the changes in the average percentile score from one year to the next were substantially larger in the target school (18.0 points) than in the other schools (where changes ranged from a decline of 14.8 points to an increase of 6.0 points, average = -.58). Unfortunately insufficient data were available to calculate effect sizes that captured the change in the target school in comparison to the peer schools.

(40.8%). Using a simple difference of proportions test the estimated effect size associated with this change was .23.[50] The results from the two designs were treated separately in the mixed model analysis that used designs as the level two measure. The results from the two designs were combined in the analysis that had site as the level two measure.

The WWC probably rejected this study using the one unit rule because the data only came from one school. However, multiple grades and teachers were involved in the study.

38) Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's *Reading Mastery*. *Journal of Behavior Assessment & Intervention in Children, 1*(1), 2–24. Study A.

This article included data from two different sites. (Results from the other site are given in number 30 above.) The analysis for this part of the study involved data obtained from two schools in the same district, situated a short distance from each other and with very similar socio-demographic characteristics. Each school had at least three teachers for each grade. One of the schools used *RM* as a core curriculum, while the other used *Open Court*. Pretest data from the DIBELS system on two pre-literacy skills (letter naming fluency and initial sound fluency) indicated no significant differences between the groups. The differences at baseline were .13 of the common standard deviation for one measure and .22 of a standard deviation for the other. With one measure the control group was higher and with the other the RM group was higher. Note that the differences were greater than the .05 limit used by the WWC, but smaller than .25 limit. Moreover, when the differences were averaged (for they went in opposite directions) the average fell within the WWC established limits. The authors used growth curve analysis to examine changes over time in nonsense word fluency (NWF) and oral word fluency lexiles, statistically controlling for differences in initial skills.

Results for NWF indicated "a widening gap in the NWF scores of students in the two groups" over time, with students in the *RM* school having slightly lower scores at mid-kindergarten (d= -.21), but at the last testing point for NWF (spring of first grade), having higher scores (d=.24). Note that for the change over time the effect size (d) = .45. At the first data point for ORF, the effect size was .46 in favor of the *RM* students and this effect persisted through the end of third grade, when the effect size in favor of the *RM* students was .42. The coefficient associated with group in the linear growth models for the analysis of ORF scores, which controlled for initial skills, was highly significant. In total, the average of these 3 effects was .44.

This article appears to have been rejected for consideration because the comparison involved data from only two schools within the same district; that is, only one school in each

[50] Caution should be used in interpreting this value because we did not have the actual panel data for the students. The appropriate test would have used data regarding the change for each student.

condition. However, there were multiple teachers at each grade and the data spanned four years of elementary school. In addition, the schools were very similar in baseline literacy skills and socio-demographic characteristics. Although the pretest differences slightly exceeded the criterion of .05 s.d. they were smaller than .25 s.d., and, when averaged, were below the criterion. Moreover, as allowed by WWC standards, multivariate analyses were used to adjust for the differences.

# Appendix E
# Summary List of Studies Identified by the WWC That Could Have Been Considered

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

2) Fredrick, L. D., Keel, M. C., & Neel, J. H. (2002). Making the most of instructional time: Teaching reading at an accelerated rate to students at risk. *Journal of Direct Instruction, 2*(1), 57–63.

3) Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *Journal of Special Education, 36*(2), 69–79; and Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education, 39*(2), 66–85.

4) Joseph, B. L. (2000). *Teacher expectations of low-SES preschool and elementary children: Implications of a research-validated instructional intervention for curriculum policy and school reform*. Dissertation Abstracts International, 65(01), 35A; and Vitale, M. & Joseph, B. (2008). Broadening the institutional value of Direct Instruction implemented in a low-SES elementary school: Implications for scale-up and school reform. *Journal of Direct Instruction, 8*(1), 1-18.

5) Marchand-Martella, N. E., Martella, R. C., Kolts, R. L., Mitchell, D., & Mitchell, C. (2006). Effects of a three-tier strategic model of intensifying instruction using a research-based core reading program in grades K–3. *Journal of Direct Instruction, 6*(1), 49–72; and Marchand-Martella, N. E., Ruby, S. F., & Martella, R. C. (2007). Intensifying reading instruction for students within a three-tier model: Standard-protocol and problem solving approaches within a response-to-intervention (RTI) system. *TEACHING Exceptional Children Plus, 3*(5). We have not included this second analysis in our compilation of results.

6) O'Brien, D. M., & Ware, A. M. (2002). Implementing research-based reading programs in the Fort Worth independent school district. *Journal of Education for Students Placed at Risk, 7*(2), 167–195. (rejected by WWC for unacceptable design) (Design A)

7) SRA/McGraw-Hill. (2005b). *Barren County elementary schools post highest reading scores ever*. Columbus, OH: The McGraw-Hill Companies.

8) SRA/McGraw-Hill. (2005d). *Delaware charter school students maintain high reading scores*. Columbus, OH: The McGraw-Hill Companies.

9) SRA/McGraw-Hill. (2005e). *Direct Instruction helps Kentucky blue ribbon school attain record reading scores*. Columbus, OH: The McGraw-Hill Companies.

10) SRA/McGraw-Hill. (2005h). *Milwaukee elementary nearly doubles reading scores*. Columbus, OH: The McGraw-Hill Companies.

11) SRA/McGraw-Hill. (2005i). *Oregon Reading First project uses Reading Mastery Plus as core reading program*. Columbus, OH: The McGraw-Hill Companies.

12) SRA/McGraw-Hill. (2005j). *Phoenix inner-city students strive toward national reading average*. Columbus, OH: The McGraw-Hill Companies. (Note this was also listed by the WWC as SRA/McGraw Hill (n.d.m)

13) SRA/McGraw-Hill. (2005l). *Reading Mastery Plus helps Colorado school achieve AYP for first time*. Columbus, OH: The McGraw-Hill Companies.

14) SRA/McGraw-Hill. (2006a). *Cleveland school keeps Reading Mastery as curriculum core*. Columbus, OH: The McGraw-Hill Companies.

15) SRA/McGraw-Hill. (2006b). *DIBELS scores advance to grade level with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

16) SRA/McGraw-Hill. (2006e). *Native American school uses Reading First grant to implement Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

17) SRA/McGraw-Hill. (2007b). *Low-performing Kentucky school on its way to high-performing with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

18) SRA/McGraw-Hill. (2007d). *Reading scores rise at Alabama elementary school with Reading Mastery Plus*. Columbus, OH: The McGraw-Hill Companies.

19) SRA/McGraw-Hill. (2007f). Title I schools in North Carolina district meet all-state reading targets with Direct Instruction. Columbus, OH: The McGraw-Hill Companies.

20) SRA/McGraw-Hill. (n.d.a) Seattle school boosts reading scores with Reading Mastery curriculum. Columbus, OH: The McGraw-Hill Companies.

21) SRA/McGraw-Hill. (n.d.b). *Anchorage school's diverse population flourishes with Direct Instruction*. Columbus, OH: The McGraw-Hill Companies.

22) SRA/McGraw-Hill. (n.d.n). "Nebraska District Outscores Peers Statewide, " pp. 14-15 in *Results with Reading Mastery*. Columbus, OH: The McGraw-Hill Companies.

23) SRA/McGraw-Hill. (n.d.p). *Success begins early at Alaskan elementary school*. Columbus, OH: The McGraw-Hill Companies.

24) Stockard, J. (2010). *The impact of Reading Mastery in kindergarten on reading achievement through the primary grades: A cohort control group design*. Eugene, OR: National Institute for Direct Instruction. (rejected by WWC for unacceptable design)

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

27) Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology, 47*, 187-214.

28) McIntyre, E., Rightmyer, E. C., & Petrosko, J. P. (2008). Scripted and non-scripted reading instructional models: Effects on the phonics and reading achievement of first-grade struggling readers. *Reading and Writing Quarterly, 24*(4), 377–407; and Rightmyer, E. C., McIntyre, E., & Petrosko, J. P. (2006). Instruction, development, and achievement of struggling primary grade readers. *Reading Research and Instruction, 45*, 209–241.

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

30) Stockard, J. (2011). Increasing reading skills in rural areas: An analysis of three school districts. *Journal of Research in Rural Education, 26*(8); and Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's Reading Mastery. Journal of Behavior Assessment & Intervention in Children, 1(1), 2–24. Study B.

31) Carlson, C. D., & Francis, D. J. (2002). Increasing the reading achievement of at-risk children through Direct Instruction: Evaluation of the Rodeo Institute for Teacher Excellence (RITE). *Journal of Education for Students Placed At Risk, 7*(2), 141-166. Reprinted in *Journal of Direct Instruction, 3*(1), 29–50.

32) Jones, C. D. (2002). *Effects of Direct Instruction programs on the phonemic awareness abilities of kindergarten students*. Dissertation Abstracts International, 63(03), 902A.

33) Mac Iver, M. A., & Kemper, E. (2002). The impact of Direct Instruction on elementary students' reading achievement in an urban school district. *Journal of Education for Students Placed at Risk, 7*(2), 197–220.

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

35) Ashworth, D. R. (1999). Effects of Direct Instruction and basal reading instruction programs on the reading achievement of second graders. *Reading Improvement, 35*(4), 150–156.

36) Green, A. K. (2010). *Comparing the efficacy of SRA* Reading Mastery *and guided reading on reading achievement in struggling readers*. Dissertation Abstracts International, 71(11A), 3969.

37) SRA/McGraw-Hill. (2009). *A report on the effects of SRA/McGraw-Hill's Reading Mastery, Signature Edition: A response to intervention solution*. DeSoto, TX: Author. (Note that there are two different designs in this study.)

38) Stockard, J., & Engelmann, K. (2010). The development of early academic success: The impact of Direct Instruction's Reading Mastery. *Journal of Behavior Assessment & Intervention in Children, 1*(1), 2–24. Study A.

Studies numbered 1 to 24 were rejected by the WWC at step 22. The remaining studies were rejected at step 3. For the mixed models (meta-analysis) reported in the companion technical report each of the 37 studies listed above was treated as a separate unit for the analysis that used design as the level 2 factor. For the analysis that used site as the second level variable, results from the following studies were combined: numbers 1 and 33, 6 and 29, 24 and 30, and the two study designs included in number 37. There were 38 level 2 units for the mixed model analysis that used design as the level 2 unit and 33 level 2 units for the mixed model analysis that used site as the level 2 unit.

Table E-1 summarizes the characteristics and results of these studies. The first column gives the study number, corresponding to the listings in Appendices C and D, and the second column gives the average effect size. The next set of columns summarize characteristics of the studies that are related to the WWC criteria and standards including

the presence of random assignment, use of multivariate statistical adjustments, a pretest-posttest control group design, the use of cohort control groups, characteristics related to the one unit rule (only two schools, only one school and only one district), and differences at pretest that exceed the WWC limits. Also included is the number of students in the study, the step at which the WWC rejected the study and the number of effects that were calculated from the study. These are the raw data that were used in the mixed model analyses described in the companion report.

Table E-1

*Characteristics of Studies Using Design as the Level 2 Measure*

| Study Number | Average Effect Size | Random Assign. | Statistical Adjust. | Pretest-Posttest | Cohort Cont. Gp.. | Two Schools | Only One School | Only One District | Not Equal at Pretest | Number of Students | Rejected at Step | Number of Effects |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.11 | No | No | No | No | Yes | No | Yes | No | 4800 | Two | 60 |
| 2 | 0.58 | No | Yes | Yes | No | No | Yes | Yes | No | 107 | Two | 8 |
| 3 | 0.31 | Yes | Yes | Yes | No | No | No | No | Yes | 256 | Two | 47 |
| 4 | 0.60 | No | No | No | Yes | No | Yes | Yes | No | 1000 | Two | 18 |
| 5 | 0.13 | No | No | Yes | No | No | Yes | No | No | 184 | Two | 5 |
| 6 | 0.14 | No | No | No | Yes | No | No | Yes | Yes | 22078 | Two | 9 |
| 7 | 0.91 | No | No | No | Yes | No | No | Yes | No | 241 | Two | 5 |
| 8 | 0.79 | No | No | No | Yes | No | Yes | Yes | No | 40 | Two | 1 |
| 9 | 0.22 | No | No | No | Yes | No | Yes | Yes | No | 220 | Two | 2 |
| 10 | 0.56 | No | No | No | Yes | No | Yes | Yes | No | 146 | Two | 1 |
| 11 | 0.73 | No | No | No | Yes | No | No | Yes | No | 300 | Two | 2 |
| 12 | 0.85 | No | No | No | Yes | No | Yes | Yes | No | 320 | Two | 1 |
| 13 | 0.63 | No | No | No | Yes | No | Yes | Yes | No | 80 | Two | 1 |
| 14 | 0.79 | No | No | No | Yes | No | Yes | Yes | No | 90 | Two | 2 |
| 15 | 0.40 | No | No | No | Yes | No | Yes | Yes | No | 200 | Two | 3 |
| 16 | 0.90 | No | No | No | Yes | No | Yes | Yes | No | 136 | Two | 4 |
| 17 | 1.04 | No | No | No | Yes | No | Yes | Yes | No | 72 | Two | 1 |
| 18 | 0.11 | No | No | No | Yes | No | Yes | Yes | No | 131 | Two | 1 |
| 19 | 0.45 | No | No | No | Yes | No | No | Yes | No | 574 | Two | 2 |
| 20 | 0.61 | No | No | No | Yes | No | Yes | Yes | No | 96 | Two | 1 |
| 21 | 0.83 | No | No | No | Yes | No | Yes | Yes | No | 118 | Two | 1 |
| 22 | 0.50 | No | No | No | Yes | No | No | Yes | No | 1232 | Two | 4 |
| 23 | 0.48 | No | No | No | Yes | No | Yes | Yes | No | 164 | Two | 5 |
| 24 | 0.45 | No | No | No | Yes | No | No | No | No | 775 | Two | 2 |
| 26 | 0.97 | No | Yes | Yes | No | No | No | Yes | No | 119 | Three | 4 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | 0.23 | No | Yes | Yes | No | Yes | No | No | Yes | 21003 | Three | 15 |
| 28 | 0.11 | No | No | Yes | No | No | No | No | No | 108 | Three | 13 |
| 29 | 0.26 | No | Yes | Yes | No | No | No | Yes | No | 22078 | Three | 12 |
| 30 | 0.57 | No | Yes | Yes | Yes | No | No | No | No | 1689 | Three | 6 |
| 31 | 0.79 | No | Yes | Yes | No | Yes | No | Yes | No | 20508 | Three | 17 |
| 32 | 0.49 | Yes | Yes | Yes | No | No | Yes | Yes | No | 36 | Three | 1 |
| 33 | 0.11 | No | Yes | Yes | No | Yes | No | Yes | No | 420 | Three | 4 |
| 34 | 2.44 | Yes | Yes | Yes | No | No | Yes | Yes | Yes | 31 | Three | 4 |
| 35 | 1.60 | No | No | Yes | Yes | No | Yes | Yes | No | 42 | Three | 1 |
| 36 | -0.53 | No | Yes | Yes | No | Yes | No | Yes | No | 66 | Three | 2 |
| 37A | 1.13 | No | No | Yes | Yes | No | Yes | Yes | No | 249 | Three | 4 |
| 38 | 0.44 | No | Yes | Yes | No | Yes | No | Yes | No | 169 | Three | 3 |
| 37B | 0.23 | No | No | Yes | No | No | Yes | Yes | No | 33 | Three | 1 |

## References[51]

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

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

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

Coughlin, C. (2014). Outcomes of Engelmann's Direct Instruction: Research Syntheses, pp. 25-54 in J. Stockard (Ed.). *The Science and Success of Engelmann's Direct Instruction*. Eugene, OR: NIFDI Press.

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

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

Horner, R.D. & Baer, D.M. (1978). Multiple-probe technique. *Journal of Applied Behavior Analysis,11*(1), 189-196.

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

Przychodzin-Havis, A. M., Marchand-Martella, N. E., Martella, R. C., Miller, D. A., Warner, L., Leonard, B., et al. (2005). An analysis of *Corrective Reading* research. *Journal of Direct Instruction, 5*(1), 37–65.

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

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

Stockard, J. (2013) *Examining the What Works Clearinghouse and Its Reviews of Direct Instruction Programs*. Eugene, OR: National Institute for Direct Instruction.

Stockard, J. & Wood, T.W. (2013). *The WWC Review Process: An Analysis of Errors in Two Recent Reports.* Eugene, OR: NIFDI Technical Report 2013-4.

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

[51] References to studies included in the WWC report are in Appendix A. Those of studies omitted from the WWC listing are in Appendix B.

What Works Clearinghouse (2012). WWC Evidence Review Protocol for Beginning Reading Interventions, Version 2.1. Retrieved from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=27, retrieved August 23, 2014.

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

What Works Clearinghouse (2013b). *WWC intervention report, Reading Mastery and beginning reading*. Washington, D.C.: Institute of Education Sciences. Retrieved December 13, 2013, from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/WWC_ReadingMastery_081208.pdf

What Works Clearinghouse (2014). WWC procedures and standards handbook (Version 3.0). Washington, D.C.: Institute of Education Sciences. Retrieved August 13, 2014 from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19