A Summary of Concerns Regarding the What Works Clearinghouse

A NIFDI White Paper

Jean Stockard, Ph.D.
Director of Research, National Institute for Direct Instruction
September 4, 2012

There are many studies that have examined the extent to which Direct Instruction (DI) curricula promote student achievement. Meta analyses of these works have found strong evidence that the programs are highly effective. For instance, in their meta-analysis of 29 comprehensive school reform models, Borman and associates found that the most evidence was available for the Direct Instruction model with "49 studies with 182 outcomes" compared to a median of four studies and 23 outcomes (Borman, et al., 2003, p. 141). DI was found to produce the strongest effects of all models examined and was one of three models that met their criteria of "strongest evidence of effectiveness," which involved replication of the outcomes "in a number of contexts, …statistically significant and positive achievement effects in studies using comparison groups or third-party comparison designs and…accumulated evidence from at least 5 third-party comparison studies" (p. 161). Other meta-analyses echo these results. For instance, Adams and Engelmann's (1995) of 37 studies found that 87% of the 173 comparisons examined favored Direct Instruction and that the average effect size across all studies was .75. White's (1988) meta-analysis of studies of Direct Instruction with special education students looked at 25 studies and found an average effect size across all comparisons of .84. Coughlin's (2011) meta-analysis was limited to 25 randomized control trials, with 95 separate comparisons, and found an average effect size of .66. Hattie (2009) summarized the results of four meta-analyses that included DI, incorporating 304 studies, 597 effects and over 42,000 students. He found that the average effect size associated with DI was .59 and noted that the positive results were "similar for regular (d [the effect size] =.99) and special education and lower ability students (d=0.86), … [and] similar for the more low-level word-attack (d=.64) and also for high-level comprehension (d=.54)" (pp. 206-207). Direct Instruction was the only curricular approach with such strong support.

Similar positive results appear in narrative reviews of specific Direct Instruction programs. Kinder, Kubina, and Marchand-Martella (2005) examined 45 studies of the use of Direct Instruction with students with disabilities and found positive effects in over 90 percent of the analyses. In 2004 Przychodzin and colleagues (Pryzhchodzin, Marchand-Martella, Martela, & Azim, 2004) reviewed 12 studies of DI math programs published since 1990 and reported that all but one showed positive results. In 2005 Przychodzin and colleagues examined 28 published studies on Corrective Reading (Przychodzin-Havis, Marchand-Martella, Martella, Miller, Warner, & Chapman, 2005) and found positive results in 26 of the 28. Similarly, Schieffer and colleagues (Schieffer, Marchand-Martella, Martella, Simonsen, & Waldron-Soler, 2002) reviewed 21 studies of *Reading Mastery* and found positive results in over two-thirds of the articles and results in favor of other programs in only three.

These analyses, which compare and contrast the results of many different studies, build on the classical notion that science is a cumulative enterprise and on the social science literature on experimental design by Campbell, Stanley and their successors (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). These writings

stress the ways in which scientific knowledge gradually accumulates and develops through testing and replications. They also assume that there can be no "perfect" experiment. They stress the importance of looking at a variety of results in a range of settings to amass knowledge. As Cook and Campbell put it, "we stress the need for *many* tests to determine whether a causal proposition has or has not withstood falsification; such determinations cannot be made on one or two failures to achieve predicted results" (1979, p. 31, emphasis in original). The vast accumulation of literature on Direct Instruction could be seen as an example of this process. Importantly, the many different tests of the curricula have consistently produced the same conclusion: the DI programs are highly effective in a wide range of settings and with many different types of students.

The What Works Clearinghouse and Direct Instruction

The What Works Clearinghouse is a federally funded program established in 2002 that evaluates educational interventions on the basis of the "rigor of research evidence" and provides summary ratings on its website. Their rating reports began to appear in 2007, but their reports on Direct Instruction curricula contrast sharply with the cumulative results in the scholarly literature. Since 2006, the WWC has published 7 reviews of Direct Instruction curriculum: English Language Learners (ELL) in 2006; Early Childhood Education (ECH), Adolescent Literacy and Beginning Reading for Corrective Reading (ARCR, BRCR) in 2007; Beginning Reading for Reading Mastery (BRRM) in 2008; Adolescent Literacy for Reading Mastery (ALRM) in 2010; and Special Needs/Learning Disabilities for Reading Mastery (LDRM) in 2012. One of the reports (ELL, September, 2006) found potentially positive effects, the second highest possible rating, on reading achievement. Two found potentially positive effects on some of the measured areas (BRCR, ALRM). Two reports indicated no discernible effects (ECH, ALCR), one report indicated that no studies met their standards for inclusion (BRRM), and the most recent report found no discernible effects on one area and potentially negative effects in three areas (LDRM). (See Table 1 for a summary.) The reports have been widely publicized through IES listserves and other means on the WWC website.

Because the findings of the WWC reports differ so strongly from the findings of the scholarly literature, NIFDI's Office of Research and Evaluation has attempted to understand why these differences appeared and if there could be any validity to their conclusions. The results of these investigations, which have spanned the last four years, suggest, very strongly, that the WWC findings are false and misleading, involving serious misinterpretations of research articles, selection criteria and other procedures that result in very limited views of the literature, lack of consistency from one review to another, questionable technical procedures, and limited transparency or adherence to the usual norms of scientific integrity. The paragraphs below briefly summarize key aspects of these conclusions. Additional details are available in the attached documents: 1) NIFDI's 2008 analysis of the WWC's review of Reading Mastery and a record of NIFDI's correspondence with the WWC over errors in that report (Stockard, 2008), 2) NIFDI's analysis of the WWC's review of Reading Mastery for Students with Learning Disabilities (Stockard & Wood, 2012), 3) an article published in 2010 regarding the WWC's fidelity implementation policies (Stockard, 2010a); and 4) an article published in 2011 that addresses the WWC's methodological criteria and standards (Stockard, 2011). It should be noted that we are far from alone in having concerns regarding the procedures of the WWC, and the attached documents include references to some of the writings that have detailed these concerns.

Misinterpretations of the Research

The most serious problems with the WWC reports undoubtedly involve misinterpretations of the research. For instance, in the 2012 report (LDRM), the WWC found two articles that met their inclusion criteria. One article (Cooke, Gibbs, Campbell, & Schalvis, 2004) compared students receiving *Reading Mastery* with those who received *Horizons*, another Direct Instruction reading program, which, as described in great detail in the article, differs from *RM* in only minor aspects. The study involved three teachers in three different schools and a total of 30 resource room students. Within each classroom reading groups were randomly assigned to receive either *Horizons* or *RM*. Thus teachers taught both programs each day. The authors found that students in both groups had growth in reading skills over time that were substantially greater than those of students in national or state norming groups. (Effect sizes of changes in Woodcock Johnson Reading scores ranged from .09 to .36, effect sizes associated with the state literacy exams ranged from .71 to .78.) Because, however, the *Horizons* students and the *RM* students had equivalent patterns of growth, the WWC report concluded that the *RM* program had "no discernible effect." The accurate conclusion, however, would have been that both Direct Instruction programs were effective.

The other study accepted for review by the WWC (Herrera, Logan, Cooker, Morris, & Lyman, 1997) involved two groups of students, both of which received *Reading Mastery* as part of the district's "usual and customary school day curriculum." In addition, students in a group of randomly selected classrooms received 45 minutes of supplemental phonemic related instruction, from their regular classroom teachers. This additional instruction involved motor activities to accompany practice of phonetic skills. The group receiving the additional instruction had significantly larger gains than those who did not have additional learning time. The WWC used these results to suggest that *Reading Mastery* could have potentially negative effects. The more logical interpretation would be that students who received additional phonics-related instruction had stronger growth. Additional details on the Cooke, et al. and Herrera, et al. studies are given in Attachment 2 (Stockard & Wood, 2012).

Such misinterpretations have also appeared in analyses of other programs. In the case of the Reading Recovery program, the misinterpretations have resulted in positive interpretations of studies when more thorough readings indicate that the articles actually indicated other results. (See Carter & Wheldall, 2008; Reynolds, Wheldall, & Madelaine, 2009; Stockard, 2008, pp. 12-14.) For instance, one of articles cited by the WWC as showing positive impacts of Reading Recovery actually reported that success rates declined over time and that by third grade, there was no difference in achievement scores, needs for retention, special education, or Chapter 1 assignments of students who had participated in *Reading Recovery* and other students. The authors explicitly concluded that *Reading Recovery* was very expensive to implement in relation to the benefits that it provided (Baenen, Bernhole, Dulaney, & Banks, 1997). Another study cited by the WWC as an example of positive effects (Iversen &Tunmer, 1993) compared the standard *RR* program to a "modified" program that included explicit instruction in phonological skills. Students in both the unmodified *Reading Recovery* program and the program that included phonological instruction eventually caught up with other students, but those in the modified program that included phonics did so much more quickly and continued to have higher levels of achievement and higher rates of learning at the end of the school year. The authors

provided an extensive discussion and additional analyses that demonstrate the fallacy involved in *Reading Recovery* about the ways in which word recognition skills develop and clearly concluded that *Reading Recovery* is not an efficient method for teaching children to read and that phonological training is superior. The WWC, however, chose to ignore these results, focusing on only the short term growth and discounting the modified program because it involved a modification. (Details on the article and the WWC's rationale for their decisions are in Stockard, 2008.)

Other Issues

In addition to misinterpreting the research, several WWC practices and policies appear to contribute to the development of inaccurate and misleading reports. The general approach used by the What Works Clearinghouse contrasts very sharply with the classic literature on the accumulation of knowledge and the literature on research design that was described briefly above. The WWC reviews emphasize "internal validity" of research design, ignoring the results of studies that are thought to deviate from so-called "ideal" practices. Although the criteria employed (and/or their application) appear to vary from one review to another, they generally involve stringent rules regarding sample selection and study design. In short, the WWC appears to be searching for the perfectly designed experiment, implicitly suggesting that a perfectly designed experiment will give the best results – a position directly contrary to that of the Campbell and Stanley tradition noted above. This search for "perfection" leads to very few studies actually being included in their reviews. For instance, in the reports on DI curricula the WWC lists over 400 separate studies that were examined, but found only seven that met their criteria. Thus, their reports are based on only a fraction of the available evidence – an approach that is directly contrary to the tradition of cumulative research noted above. (Table 1 provides details on the numbers of studies examined for each DI report.)

Stockard (2008) describes a number of errors in the selection of studies in the Beginning Reading report on *Reading Mastery*, and Reynolds and associates (2009) and Carter and Wheldall (2008) describe similar issues with the analysis of Reading Recovery. One WWC selection criterion appears to disproportionately impact the Direct Instruction programs, for the analyses are limited to studies that appeared within 20 years of the review. Because the literature regarding Direct Instruction is so well established, this limitation automatically excludes a large number of studies from review. We know of no other curriculum that has been negatively affected by this limitation. The WWC also requires that studies include pretests of students to determine equivalency of groups. This requirement appears to eliminate virtually all large scale studies that occur in real life settings, most of which use various well established statistical techniques to ensure equivalence in comparisons.

Another problematic issue is a lack of consistency from one review to another. The discussion of the report on Reading Mastery for Beginning Readers (Stockard, 2008) includes several examples of this problem, such as acceptance of a study as meeting inclusion criteria in one review but rejection of the same study in another review. As another example, a large, comprehensive analysis of Direct Instruction (Carlson & Francis, 2002) was rejected for inclusion in the BRRM review because teachers were trained in implementing the program and in managing students' behavior, standard elements of the Direct Instruction approach. The WWC determined that this involved an alteration of *Reading*

Mastery (see Stockard, 2008). Yet, the Herrera, et al study included in LDRM could also be seen as involving an intervention that was not clear or an alteration of RM, for students received both RM and the additional instruction. Wheldall (2012) points to the recent WWC review of Open Court (OC) as yet another example. The analysis was based on one study (out of 58 examined) and claimed that there were potentially positive effects of the program on comprehension for adolescent readers. However, the selected study had an effect size of only .15, which, as Wheldall put it, is "well below the WWC's own usual threshold of .25." Interestingly, the WWC report rejected for inclusion at least one large study that compared OC with DI programs and found, as would be expected, significantly positive results in favor of DI. This study (Stockard, 2010b) was excluded "because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent." A close reading of the article shows that it compared growth in achievement from first grade (when the DI students had lower levels of reading skills) through fifth grade (when the DI students had higher levels) of approximately 4500 students in the Baltimore City School System. Statistical techniques standard to the social sciences adjusted for initial differences in achievement. (The DI students had lower average scores in first grade.) With or without these adjustments, however, the DI students had higher scores in fifth grade than the OC students. Wheldall reports a similar story of a randomized control trial that was excluded from the review of Reading Recovery (Wheldall, 2012; Carter & Wheldall, 2008), although the study had been widely praised in the literature, amassing 160 citations since publication and, upon close inspection clearly should have met the inclusion criteria.

A third type of concern involves technical elements of the review process. One example involves the procedures (or lack of such) used to handle low fidelity of implementation. As explained in Stockard (2010a), the WWC criteria can result in higher ratings given to ineffective programs and lower ratings given to more successful programs. Others have suggested that the measures utilized by the WWC are not well grounded in the theoretical or empirical literature, use invalid and overly broad indicators, and thus can lead to misleading results (Reynolds, et al., 2009). In an extensive discussion, Stockard (2011) used the classic literature on research design to challenge the narrow restrictions on acceptable research designs used by the WWC and showed how many of the case study reports of implementations within schools that are routinely rejected for inclusion by the WWC can, in fact, produce analyses that meet the Campbell and Stanley criteria for both internal and external validity.

Finally, the WWC appears to be resistant to using peer review procedures that are a standard part of the scientific process or to altering mistaken reports. Stockard (2008) includes an extensive discussion of these issues. It also includes details of NIFDI's attempts to communicate with the WWC regarding errors in their reviews, problems with their procedures, and the responses that were received. In subsequent communications (available on request), NIFDI asked that we be contacted to provide feedback on reports before they are posted on the web. Such review would help counter the possibility of misinterpretations of the research. This request was, however, denied. The WWC has recently established a "quality review" procedure, and it is hoped that this procedure may indicate more openness to review and checks on the quality of reports. NIFDI has used this procedure to ask for a review of the 2012 report on Reading Mastery and Students with Learning Disabilities, and the WWC has indicated that it will do so. They refused, however, to

remove the 2012 report from the website while the review is being conducted and would not give an estimate of the time required to complete the analysis.

References

Adams, G.L. & Engelmann, S. (1995). *Research on Direct Instruction: 25 Years Beyond Distar.* Seattle, WA: Educational Achievement Systems.

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

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

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

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

Carter, M. & Wheldall (2008). Why can't a teacher be more like a scientist? Science, pseudoscience and the art of teaching. *Australaisian Journal of Special Education, 32*, 5-21.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Coughlin, C. (2011). *Research on the Effectiveness of Direct Instruction Programs: An Updated Meta-Analysis*, Poster presented at the Annual Meetings of the Association for Behavior Analysis International, May, 2011, NIFDI Technical Report 2011-4.

Hattie, John A.C. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. London and New York: Routledge.

Iverson, S.. & Tunmer, W.E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology, 85,* 112-125.

Kinder, D., Kubina, R., & Marchand-Martella, N. (2005), Special education and Direct Instruction: An effective combination. *Journal of Direct Instruction*, 5 (1), 1-36. Also distributed by SRA/McGraw-Hill as *Special Education and Direct Instruction: An Effective Combination.*

Przychodzin, A.M., Marchand-Martella, N., Martella, R.C., & Azim, D. (2004). Direct Instruction Mathematics Programs: An Overview and Research Summary. *Journal of Direct Instruction, 4 (1)*, 53-84.

Przychodzin-Havis, A.M., Marchand-Martella, N., Martella, R.C., Miller, D.A., Warner, B.L., & Chapman, S. (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. (2008). The What Works Clearinghouse Beginning Reading Reports and Rating of *Reading Mastery*: An Evaluation and Comment

Stockard, J. (2010a). An Analysis of the Fidelity Implementation Policies of the What Works Clearinghouse, *Current Issues in Education, 13* (4).

Stockard, J. (2010b). Promoting reading achievement and countering the "fourth-grade slump": The impact of direct instruction on reading achievement in fifth grade. *Journal of Education for Students Placed at Risk, 15*(3), 218–240.

Stockard, J. (2011). Merging the Accountability and Scientific Research Requirements of the No Child Left Behind Act: Using Cohort Control Groups. *Quality and Quantity: International Journal of Methodology*, available on-line, December, 2011.

Stockard, J. & Wood, T. (2012). *Reading Mastery and Learning Disabled Students: A Comment on the What Works Clearinghouse Review*. Eugene, Oregon: National Institute for Direct Instruction.

Whedall, K. (2012). What's Wrong with What Works (www.kevinwheldall.com), downloaded September 1, 2012.

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

Table 1
WWC Reviews of DI Curricula

Report Date	Topic/ Protocol	Program	Studies Reviewed	Met Standards	Met with reservations	Outcome
September, 2006	English Language Learners	RM	1	1	0	Potentially positive effects on reading achievement
July, 2007	Beginning Reading	CR	25	1	0	Potentially positive effects on fluency and alphabetics; no discernable effects on comprehension
August, 2010	Adolescent Literacy	RM	175	1	1	Potentially positive effects on fluency, no discernable effects on comprehension
May, 2007	Early Childhood Education	DI	6	0	1	No discernable effects in mathematics, oral language, cognition or print knowledge
July, 2007	Adolescent Literacy	CR	129	1	0	No discernable effects
August, 2008	Beginning Reading	RM	61	0	0	No studies met standards
July, 2012	Special Needs/ LD	RM	17	2	0	No discernable effects on comprehension, potentially negative effects on alphabetics, fluency, and writing