Published in final edited form as:

Except Child. 2019 January ; 85(2): 229–247. doi:10.1177/0014402918802801.

Using Moderator Analysis to Identify the First-Grade Children Who Benefit More and Less from a Reading Comprehension Program: A Step Toward Aptitude-by-Treatment Interaction

Douglas Fuchs[1], **Devin M. Kearns**[2], **Lynn S. Fuchs**[1], **Amy M. Elleman**[3], **Jennifer K. Gilbert**[1], **Samuel Patton**[1], **Peng Peng**[4], and **Donald L. Compton**[5]

[1]Vanderbilt University

[2]University of Connecticut

[3]Middle Tennessee State University

[4]University of Nebraska

[5]Florida State University

Abstract

Because of the importance of teaching reading comprehension to struggling young readers and the infrequency with which it has been implemented and evaluated, we designed a comprehensive first-grade reading comprehension program. We conducted a component analysis of the program's decoding/fluency (DF) and reading comprehension (COMP) dimensions, creating DF and DF +COMP treatments to parse the value of COMP. Students ($N = 125$) were randomly assigned to the 2 active treatments and controls. Treatment children were tutored 3 times per week for 21 weeks in 45-min sessions. Children in DF and DF+COMP together performed more strongly than controls on word reading and comprehension. However, pretreatment word reading appeared to moderate these results such that children with weaker beginning word reading across the treatments outperformed similarly low-performing controls to a significantly greater extent than treatment children with stronger beginning word reading outperformed comparable controls. DF +COMP children did not perform better than DF children. Study limitations and implications for research and practice are discussed.

Many weak readers in the early elementary grades continue to struggle as students throughout their educational careers (Foorman, Francis, Shaywitz, Shaywitz, & Fletcher, 1997). Lasting reading problems not only depress achievement across academic domains but are associated with undesirable post-school outcomes like low income and poor health (Miller, McCardle, & Hernandez, 2010). That said, early signs of reading difficulty need not sentence children to chronic school failure and life-long misery. Proper early instruction can reduce the number of children with persistent reading problems (Torgesen, Wagner, & Rashotte, 1999).

Send inquiries to Doug Fuchs (doug.fuchs@vanderbilt.edu).

Comprehension Instruction in the Early Grades

Importance.

Conventional reading instruction in kindergarten and first and second grades focuses on word reading (Case et al., 2014). Research supports this emphasis partly because instruction in alphabetic skills and word-level reading can enhance comprehension (Vadasy & Sanders, 2011). Such an effect may be explained by Perfetti and Hart's (2001) lexical quality hypothesis, which describes word identification as word knowledge that strengthens or weakens depending on the precision of a reader's orthographic, phonological, and morphological representations and word meanings. Implicit is a bi-directional relation between so-called bottom-up and top-down cognitive processes. Nonetheless, many believe early and explicit comprehension instruction is also necessary to improve later reading because comprehension quickly supplants word reading as a pivotal skill (Catts, Fey, Thomblin, & Zhang, 2002).

Despite a growing consensus on the importance of early comprehension instruction, there is little accord on what it should look like. This is in part because of the complex nature of reading comprehension. Reading with understanding depends on cognitive processes like attention, working memory, reasoning, and inferential thinking; on sensitivity to the structure of language; on background knowledge and vocabulary development; on motivation; on the use of strategies like self-monitoring; and of course on word reading. It is understandably difficult for program developers and practitioners to identify which cognitive processes, knowledge sets, and reading comprehension strategies should be part of comprehension instruction, let alone how to fashion these components into a coherent and manageable whole and to combine it with word-level instruction.

Strategy instruction.

Such uncertainty notwithstanding, strategy instruction has emerged over the past two decades as a favored approach among many educators. What distinguishes it from skills-based methods is that it encourages students to act with intentionality; to consciously apply strategies that help them read for understanding.

After reviewing the relevant research in grades K-3, a panel of experts found strong support for teaching strategies (Shanahan et al., 2010). Among 13 investigations, all but one showed that strategy instruction improved reading comprehension. However, the panel acknowledged that most of the reviewed studies had been conducted with older children; none focused exclusively on first grade; and just three included first-grade students with older students in their study samples (i.e., Brown, Pressley, Van Meter, & Schuder, 1996; McGee & Johnson, 2003; Morrow, Rand, & Young, 1997). Only Brown and colleagues attempted to teach multiple strategies.

In a subsequent review of reading programs for at-risk first graders, Gersten, Newman-Gonchar, Haymond, and Dimino (2017) found seven additional studies that included a comprehension component. The comprehension component in these studies consisted of learning discrete comprehension skills, reading connected text, and answering questions. In none of the reviewed programs were students taught to use strategies to improve their comprehension.

The Nashville Early Reading Program

Development.

Because of the apparent importance of teaching comprehension explicitly to young at-risk readers, and the infrequency with which it has been implemented and carefully evaluated, we designed what we considered was a developmentally appropriate first-grade comprehension program with activities reflecting the multidimensionality of the construct. More specifically, our program (The Nashville Early Reading Program; D. Fuchs et al., 2015) emphasized phonological awareness, word reading, fluency-building, and multiple evidence-based strategies like identifying a story's implicit structure (Morrow, 1984), retelling and summarizing texts (D. Fuchs, Fuchs, Mathes, & Simmons, 1997; Wanzek, Wexler, Vaughn, & Ciullo, 2010), and inference-making (Elleman, 2017). It also aimed to strengthen word knowledge (Elleman, Lindo, Morphy, &, Compton, 2009) and sentence-level comprehension (Scott, 2009) and to teach students to use strategies to answer specific kinds of comprehension questions (Raphael, 1984).

Evaluation.

We explored the efficacy of The Nashville Early Reading Program with two related purposes in mind. First, we wanted to know the value of its reading comprehension component. Towards this end, we conducted a component analysis of the program's decoding/fluency (DF) dimension and reading comprehension (COMP) dimension. That is, we contrasted a DF treatment condition against a DF+COMP condition to parse the added value of COMP. We also compared the two conditions when combined against controls. Students in the two treatment conditions were tutored three times per week for 21 weeks, with each session lasting 45 min. Thus, the first graders in DF obtained 45 min of DF instruction; those in DF +COMP got 30 min and 15 min of DF and COMP instruction, respectively, a point to which we return.

Our second purpose was to explore whether the efficacy of the DF and DF+COMP conditions interacted with our sample's pretreatment word reading performance. We had two expectations in this regard. First, children with *weaker* pretreatment word reading might benefit more from the DF condition than children with stronger pretreatment word reading because they would arguably be in greater need of such a focus. Second, those with *stronger* word reading might respond more positively to the DF+COMP condition because word reading would not be as much a "bottleneck" for them as for the poorer readers; that is, the stronger readers would have more cognitive resources available to read with understanding.

Put differently, our second study purpose was to use pretreatment word reading as a moderator to explore for whom our program, in its less inclusive (DF) and more inclusive (DF+COMP) variants, was beneficial. Whereas others have also used moderation analysis for this purpose (e.g., Connor, Morrison, & Katch, 2004; Lang et al., 2009; Schunemann, Sporer, & Brunstein, 2013), such exploration is infrequent. This is unfortunate because moderation analysis can be a means by which interventionists better understand the nature and effects of their interventions. As discussed below, our own moderation analysis was not without problems. Our biggest concern was our small sample size and inadequate statistical power to detect reliable interaction terms (Aguinis, 1995). Although we avoided several factors that contribute to low statistical power (e.g., restriction of range, measurement error, and multicollinearity), such precautions did little to compensate for the fact that we were principally interested in exploring main effects of The Nashville Early Reading Project. Thus, we view our study's second purpose as heuristic. We graphed all moderator effects, statistically significant and not, to document patterns of effects that might inform future, more adequately powered research.

Method

Participants

To make the most of our resources and limit demands on our school-based partners, we simultaneously recruited participants for this first-grade reading study and a second, smaller first-grade reading study. We used the same eligibility criteria and screening procedures for both studies, and randomly assigned eligible students to them at the end of the screening phase. Because we cannot separate the children recruited and screened for the two projects, we first explain the common recruitment and screening procedures. Then, we describe this study.

Identifying participants and assigning them to the reading and math studies.—We recruited students from 73 first-grade classrooms and 13 schools in the Metro-Nashville Public Schools. At our request, teachers nominated their lowest readers for both studies (n =532). We screened those for whom we obtained consents (n = 389) on reading measures (see below), creating a factor score for each child. On this score, we rank ordered the 389 children and eliminated the top 150 (nearly 40%). The remaining 239 students were tested on two subtests of the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999). The 10 children with a T score below 37 on both subtests were removed from the sample. So, too, were six English Language Learners (ELL) who did not have "Early Advanced" or greater proficiency at the end of kindergarten on the district's ELL assessment. This resulted in 223 study-eligible children who were randomly assigned to the larger reading study (n = 132), a smaller reading study (n = 51), and a wait list (n = 40) serving the two studies.

Assigning participants to study conditions.—The 132 students in the present reading study were then randomly assigned to DF, DF+COMP, or controls. A low rate of attrition (5.3%) resulted in a final sample of 125: 43, 40, and 42 children were in DF, DF +COMP, and control groups, respectively. One-way ANOVAs indicated no reliable differences between attriting and remaining students on the screening measures except that remaining students had a lower WASI Matrix Reasoning T-scores, $F(1, 130) = 9.86$, $p = .002$.

Table 1 provides student demographics. Nearly 70% of the DF+COMP group, but only 50% of controls, were male. Between 50% and 60% of each group were children of color, and a higher proportion in each group received free or reduced lunch. Between 15% and 19% of the children were ELL. Despite several of the between-group, demographic differences reported here and in the table, the three groups were comparable on sex, χ^2 (2) = 3.13, p = .209; race, χ^2 (2) = 0.94, p = .626; ELL status, χ^2 (2) = 0.35, p = .838; and whether they received free/reduced lunch, χ^2 (2) = 1.22, p = .544. Neither IEP nor retention designations were subjected to a χ^2 test because of near-zero cell counts. DF, DF+COMP, and control students' Full Scale WASI scores were 89.73 (*SD* = 9.06), 88.95 (*SD* = 8.99), and 90.91 (*SD* = 11.98), respectively. Corresponding percentiles are 27.4, 26.3, and 30.2.

There were 63 teachers of the 125 children in the final sample. Many taught students in all three study conditions. On average, they had been teaching 15 years, with 7 years in their current position. Most were female (96.8%), white non-Hispanic (82.5%), and had a master's degree (58.7%). A majority (52.4%) had 0–3 credit hours of special education training.

Measures

We chose measures to accomplish two objectives: screen children into the study and determine tutoring outcomes. Our *screening measures* documented children's alphabetic knowledge and word- and non-word reading skills. Alphabetic measures included a Rapid Letter Naming Test (D. Fuchs et al., 2001) and a Rapid Sound Naming Test (D. Fuchs et al., 2001). Word Reading was assessed with the Sight Word Efficiency subtest of the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999), Word Identification Fluency A (WIF A; L. Fuchs, Fuchs, & Compton, 2004), and the Word Identification subtest of the Woodcock Reading Mastery Test – Revised (WRMT; Woodcock, 1998). Non-word reading was indexed by TOWRE-Phonemic Decoding Efficiency (Torgesen et al., 1999) and WMRT-Word Attack (Woodcock, 1998).

Our *outcome measures* included the word- and nonword-reading tests just mentioned except that we substituted Word Identification Fluency B (WIF B; Zumeta, Compton, & Fuchs, 2012) for WIF A. Two comprehension measures—the Passage Comprehension subtest of the WRMT (Woodcock, 1998) and the Reading Comprehension subtest of the Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001)—were also administered. Finally, we developed a DF placement test as neither screen or outcome measure, but to help tutors determine for each child the book level most appropriate to begin instruction. Appendix A provides description of all the screening, outcome, and other measures.

Tutoring

Overview.

Tutors were masters and doctoral students, post-doctoral fellows, and full-time staff. They tutored children 1:1 outside the classroom, three times per week, for 21 weeks (63 sessions). Each DF and DF+COMP session was designed to last 45 min. For the DF+COMP condition, the order of DF instruction and COMP instruction was counterbalanced so that in odd number weeks students received DF instruction first; in even number weeks, after COMP instruction. The tutoring activities were scripted and included correction procedures for incorrect responses. Much of the DF and COMP instruction was delivered as game-like activity to hold students' attention and provide many opportunities to practice targeted skills. Tutors awarded points for appropriate behavior, correct work, and effort.

Decoding/Fluency (DF).—The DF curriculum included 10 short narrative stories and scripted lessons. The books and lessons were ordered by reading difficulty. Children's scores on our DF placement test determined their first book and initial lesson. They repeated or skipped lessons based on performance. Each lesson had six parts: sight words, sound-symbol correspondence, decodable words, spelling, fluency building, and supplemental activities.

In each lesson, tutors read 12 *Sight Words* to the students. The students then read and re-read the words in various game-like contexts. Sight-word reading was also emphasized with "Sight Word Challenge." Students read as many words as they could in 1 min from a list of 500 high-frequency sight words, ordered by their frequency in the Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995). *Sound-Symbol Activities* introduced single phoneme-grapheme correspondences, digraphs, vowel-consonant-E, phonograms, past tense and letters or digraphs representing multiple phonemes. Tutors taught *Decodable Words* by tapping and sounding out words that were displayed such that single letter graphemes had dots under them and digraphs and vowel-consonant-E patterns had lines beneath them. The tutor pointed to the first grapheme in a word; pronounced its sound and the remaining sounds in the word (while tapping the dots or lines); then read the word slowly; and finally read it naturally. The student then attempted to do this same.

For *Spelling*, students were given a pile of small tiles, each displaying a single sound-symbol correspondence. They used the tiles to spell the same decodable words in the decodable word activity. For *Fluency Building*, the tutor presented a passage from one of the 10 stories and previewed difficult words. The tutor and child next read the passage together, sentence by sentence, and then the whole text. Finally, the student read the passage twice independently and aloud. The tutor timed the student's reading and encouraged faster reading on the second trial. *Supplemental Activity* was the final DF lesson component, lasting 15 min. Its purpose was to engage children in games designed to reinforce the phonological awareness and decoding skills introduced earlier in the lesson.

Reading Comprehension (COMP).—The COMP part of each DF+COMP lesson included four activities: vocabulary, big idea, wrapping up, and supplemental activities. *Vocabulary* instruction involved the tutor and child previewing the meanings of unusual or otherwise difficult words prior to reading the story for the day. The tutor read the story, and asked the student to retell it in his or her words, providing feedback if the retelling was inaccurate. Finally, students constructed a *Big Idea* for the story by identifying: "the most important person" and the "most important thing that happened."

In *Wrapping Up*, students answered four question types based on the day's story: factual questions, main-ideas questions, inference questions, and vocabulary questions. The questions were presented in multiple choice, cloze, and free-response formats. **S**tudents completed *Supplemental Activities* at the end of each lesson, as time allowed. They were designed to improve skills and knowledge associated with discourse-level processing (e.g., reasoning and inferencing skills, background knowledge, classifying objects and concepts) and sentence-level processing (e.g., using conjunctions resolving ambiguous pronouns).

Procedures

Twelve graduate students and two postdoctoral students (hereafter “research assistants,” or RAs) and two project staff conducted testing and tutoring. Pretreatment testing occurred between late August and early October. Tutoring began in late October. Posttreatment testing was conducted between late April and mid-May.

Test training and testing fidelity.

RAs were trained to administer tests in five sessions at pretreatment and four sessions at posttreatment. Prior to pre- and posttreatment testing they had to achieve 90% compliance with administration and scoring procedures. All testing was audio-recorded. To explore inter-rater agreement on the fidelity with which tests were administered, 15% of the audiotaped sessions, chosen at random (blocking by pre- and posttreatment testing and study condition), were rescored by other project staff. Agreement between testers and project staff exceeded 95% for all tests except WASI Vocabulary (88.45%). All testers were unfamiliar to the students they tested and unaware of students’ study condition. Testing was conducted 1:1.

Tutor training and tutoring fidelity.

RAs were trained as tutors in two half-day sessions. This training included an overview and rationale for each intervention component; hands-on demonstration of all instructional activities; and role-playing and practice with feedback. After each session, the RAs practiced implementing the tutoring protocols (for both DF and DF+COMP) with another RA for at least 6 hours. They were then required to conduct two lessons (one DF, one DF+COMP) with at least 90% adherence. Failing this, they engaged in more practice and their fidelity was evaluated again. Once tutoring commenced, the RAs and project staff participated in weekly 1-hour meetings to discuss upcoming lessons and unforeseen difficulties such as children’s problem behavior.

Project staff observed each RA in four tutoring sessions spaced evenly across the 21-week intervention period. RAs’ adherence to the tutoring protocols was scored with a 191-item checklist that included all required tutoring activities, organized by treatment components. Staff marked checklist items as performed correctly, performed incorrectly, or not observed. Scores were calculated by dividing the number of items conducted correctly by those observed to be correctly and incorrectly implemented. Fidelity was consistently strong (>90%) irrespective of time point and treatment condition.

Data Analysis

Calculation of factor/composite scores.

To create more robust representations of domains of interest, we calculated either factor or composite scores to index Word Reading, Non-Word Reading, and Reading Comprehension performance at pre- and posttreatment. A factor score was computed for Word Reading based on TOWRE - Sight Word Efficiency, Word Identification Fluency, and WRMT - Word Identification using Stata's *factor* command, which specifies the principal factor method and regression-based factor score calculation. Composite scores were calculated for (a) Non-Word Reading based on TOWRE-Phonemic Decoding Efficiency and WRMT – Word Attack, and (b) Reading Comprehension based on the ITBS – Reading Comprehension and WRMT – Passage Comprehension. For Word Reading and Comprehension, the measures were weighted equally in a regression-based factor score. There were no reliable differences among study conditions on pretreatment factor or composite scores (*p*s = .36 to .90).

Model set up.

To examine the efficacy of the two treatment conditions and account for school and classroom clustering, we used multilevel models. Prior research (Conner et al., 2004) and theory (Perfetti & Hart, 2001) suggest that the efficacy of reading interventions differ depending on students' pretreatment word reading. Thus, the final models are described as "moderation" models.

The outcomes of interest were posttreatment Word Reading, Non-Word Reading, and Reading Comprehension. Our models had random and fixed components. Because the study sample came from different schools and classrooms, we used three-level models in all analyses in which students (at Level 1, $n = 125$) were nested within classrooms (at Level 2, $j = 56$), which were nested within schools (at Level 3, $k = 26$). Residual terms for the intercept were allowed at Levels 2 and 3. Independent residual terms were permitted for the three conditions (Roberts & Roberts, 2005) and were retained where likelihood ratio tests indicated such terms were necessary compared to a mixed χ^2 distribution (Stram & Lee, 1994).

For the fixed effects, we examined treatment effects, moderated treatment effects with pretreatment Word Reading as a moderator, and two covariates in each model: (a) the relevant pretreatment score (e.g., controlling for pretreatment Non-Word Reading factor score when evaluating the posttreatment Non-Word Reading factor score), and (b) pretreatment Word Reading (the moderator of interest). Treatment effects were estimated by two orthogonal contrasts. The first compared both treatments together against controls (TRT_C: DF+COMP = .333, DF = .333, Control = −.667). The second compared DF +COMP to DF (DFCOMP_ DF: DF+COMP = .5, DF = −.5, Control = 0; Cohen, Cohen, West, & Aiken, 2003). Due to our relatively small sample size, we used restricted maximum likelihood estimation with the Kenward-Roger degrees of freedom adjustment (McNeish & Stapleton, 2016). The following equation represents our final multilevel (assuming homogeneous residuals across conditions at level 1):

$$\text{Posttreatment score}_{ijk} = \gamma_0 + \gamma_1 *(\text{pretreatment score}_{ijk}) + \gamma_2*(\text{TRT\_C}_{ijk}) + \gamma_3*(\text{DFCOMP\_DF}_{ijk}) + \gamma_4*(\text{pretreatment Word Reading}_{ijk}) + \gamma_5*(\text{pretreatment Word Reading}_{ijk})*(\text{TRT\_C}_{ijk}) + \gamma_6*(\text{pretreatment Word Reading}_{ijk})*(\text{DFCOMP\_DF}_{ijk}) + u_{0k} + r_{0jk} + e_{ijk}.$$

In this equation, posttreatment score$_{ijk}$ is a posttreatment factor/composite score for a given outcome (e.g., Word Reading factor score) for student *i* in classroom *j* at school *k*, and γ_0 is the grand mean intercept, $\gamma_1 - \gamma_6$ are the slopes for the fixed effects controlling for the other parameters, with the latter two representing moderated treatment effects, u_{0k} is a school-specific intercept residual, r_{0jk} is a classroom-specific intercept residual, and e_{ijk} is an error residual. Typically, significant interactions are followed by investigation of regions of significance. However, we did not calculate regions of significance because specific values of a factor/composite score are not particularly informative or generalizable and because we chose to focus on patterns of effects for all interaction terms regardless of statistical significance to generate hypotheses for future research, as indicated earlier.

Results

Pre- and Posttreatment Performance on Reading Measures

Table 2 shows means (and *SD*s) and percentile scores of the study groups' pre- and posttreatment performance on word reading, non-word reading, and comprehension measures. On most word and non-word reading measures, and on one of two comprehension tests, our sample's pretreatment performance was near the 50th percentile. This suggests that in spite of concerted efforts to identify poor readers we recruited average readers. However, percentiles notwithstanding, we believe this is an incorrect conclusion. Most measures in Table 2 are associated with strong floor effects (cf. D. Fuchs et al., 2018), reflecting that they were *not* constructed to explore the reading skill of weak beginning readers nor to differentiate among them for instructional purposes. By contrast, WIF A (L. Fuchs et al., 2004) assesses first-grade reading behavior more broadly and is more capable of distinguishing among children at the low-end of the distribution (see Appendix A). Percentiles associated with the mean WIF A pretreatment scores of DF, DF+COMP, and control groups, respectively, were 6.9, 5.9, and 5.9.

Table 2 also shows that many posttreatment means of the DF and DF+COMP groups exceeded corresponding pretreatment means, and that these gains are accompanied by positive percentile changes. In some cases, dramatically so. For example, the DF group improved its percentile ranking from 24.9 to 43.6 on ITBS Reading Comprehension; comparable change for DF+COMP was 26.7 to 42.1. Such apparent improvements were not observed among control children. Indeed, on several measures, their percentile rankings decreased over the same period.

Model Results

Models for word reading, non-word reading and reading comprehension outcomes were run using the *quietly* command to hide model results while checking for normality and homoscedasticity of level 1 residuals. Models with and without condition-specific level-1 residuals were compared via the likelihood ratio test compared to a mixed χ^2 distribution (Stram & Lee, 1994). Only the non-word outcome model required condition-specific residuals. Level-1 residuals from all final models met normality and homoscedasticity assumptions.

Main effects.

Table 3 shows fixed and random effects for each of the three outcome models. Statistically significant differences were identified for the two active treatments combined (DF and DF +COMP) versus controls on all outcomes: word reading, $p < .001$; non-word reading, $p < .001$; and reading comprehension, $p = .002$. Respective effect sizes were 0.77, 1.44, and 0.37. As notable as these outcomes may be, two additional findings add important nuance to them. First, no statistically significant effects were obtained for the DF versus DF+COMP contrast. Second, as elaborated next, treatment effects on word and non-word reading and reading comprehension tended to be moderated by children's pretreatment word reading.

Moderation effects.

After controlling for pretreatment performance and treatment main effects, moderation effects for pretreatment word reading on posttreatment *word reading* were $\hat{\gamma}_5 = -0.37$, $SE = 0.16$, $p = .020$ for the DF and DF+COMP combined treatment versus control; and $\hat{\gamma}_6 = -0.36$, $SE = 0.19$, $p = .055$ for DF vs. DF+COMP. Figure 1 shows that the combined treatments had statistically significantly greater impact than the control condition at the lower end of the word reading distribution (Figure 1a). The contrast between the two treatments reveals a marginally significant disordinal interaction, or crossover effect, where DF+COMP appears to have been more effective than DF at the lower end of the distribution but DF was more effective at the upper end (Figure 1b).

With pretreatment scores and treatment main effects controlled, pretreatment word reading moderation effects on posttreatment *non-word reading* were $\hat{\gamma}_5 = -0.26$, $SE = 0.16$, $p = .111$ for the combined treatments versus control effect; and $\hat{\gamma}_6 = -0.31$, $SE = 0.23$, $p = .186$ for DF versus DF+COMP. Figure 2 suggests the combined treatments had a greater effect than controls at the lower end of the word reading distribution than at the higher end (Figure 2a). The comparison between the two treatments suggests a (non-significant) disordinal interaction whereby DF+COMP was more effective at the lower end of the distribution and DF was more effective at the upper end (Figure 2b).

For posttreatment *reading comprehension*, pretreatment word reading moderation effects were $\hat{\gamma}_5 = -0.28$, $SE = 0.16$, $p = .086$ for treatment versus control, and $\hat{\gamma}_6 = -0.17$, $SE = 0.20$, $p = .371$ for DF versus DF+COMP. These results were obtained after controlling for pretreatment word reading and treatment main effects. Figure 3a shows that the combined treatments may be more effective at the lower end of the word-reading distribution than at the higher end. The comparison between the two treatments suggests DF+COMP may be more effective than DF at the lower end of the distribution; with DF again more effective at the upper end (Figure 3b). These differences, however, appear minimal, as illustrated by the nearly overlapping slopes in the figure. As mentioned in our introduction, we provide these contrasts as heuristic.

Discussion

Efficacy of the Nashville Early Reading Program

Treatment versus controls.—The general purpose of this randomized control trial was to explore the efficacy of the Nashville Early Reading Program on the word reading, non-word reading, and reading comprehension of first graders with poor reading skills. Following a 21-week intervention, the combined performances of the children in the two active treatment conditions, DF and DF+COMP, were statistically significantly stronger than controls on all outcomes including reading comprehension. Moderate-to-strong effect sizes accompanied these findings.

We also calculated the percentage of DF and DF+COMP children whose reading was “normalized” during treatment; that is, the percentage of children whose posttreatment scores surpassed a standard score of 90, or a percentile equivalent of 25. We did this to determine how many DF and DF+COMP children improved their reading in more “real-world” terms. We recognize that the 25th percentile denotes below-average performance, and that some will see it as a low-bar estimate of normalization.

We relied on WIF B, a validated measure of first-grade reading (see Appendix A), and a local normative sample of first graders who were tested on this measure during alternate weeks for 20 weeks from the start of the school year (Zumeta, Compton, & Fuchs, 2012; Appendix A). We first identified the raw score equivalent of the 25th percentile of the normative sample in the first and last week that WIF B was administered. We then found the same raw scores in the distributions of combined DF and DF+COMP pretreatment and posttreatment scores. From this we calculated a percentage of treated children with scores exceeding the 25th percentile prior to and immediately following the intervention. We determined the percentage of normalized controls in our sample in the same manner. The combined percentage of DF and DF+COMP children with normalized scores on WIF B was 6.2 (5 of 81 children) at pretreatment and 67.1 (51 of 76 children) at posttreatment. For controls, comparable pre- and posttreatment percentages were 2.3 (1 of 43 children) and 31.6 (12 of 38 children). These data, together with our significant main effects, suggest a relatively strong treatment.

DF versus DF+COMP.—Determining the value of a reading comprehension program component was also part of our evaluation. We constructed what we considered a developmentally appropriate, comprehensive instructional approach (i.e., DF+COMP) that involved phonological awareness, word reading, fluency-building, and multiple evidence-based strategies like identifying a story’s implicit structure, retelling and summarizing texts, and inference-making. We contrasted its effects against DF, an approach that focused on phonological awareness, word reading, and fluency-building. Instruction in both conditions was delivered 1:1 by well-trained tutors with fidelity. Children in DF+COMP did no better on the reading comprehension outcome, or on word-and non-word reading outcomes, than members of the DF-only group. There are many plausible explanations of this disappointing result. We offer three.

First, although our intent was to produce a developmentally appropriate comprehension program, one could argue we inadvertently produced its opposite. Requiring young at-risk children to apply cognitively taxing strategies (cf. Perfetti, 2007; Kendeou et al., 2011) as they tried to decode words may simply have been an impossible expectation, let alone an unfair and unproductive one. A second explanation does not question the appropriateness of our strategies-oriented approach, but rather how we operationalized it. Specifically, we may have tried to accomplish too much, allocating only 15 min of each 45-min DF+COMP session to comprehension instruction. Although these 15 min may seem reasonable, our tutors said they had too little time to teach too many strategies and feared their teaching suffered as a result.

A third explanation is that our two distal reading comprehension measures were insensitive to treatment aims and inadequate instantiations of the construct (cf. Keenan, Betjeman, & Olson, 2008). The ITBS seems more a test of decoding than comprehension, which raises the possibility that our comprehension composite registered word-reading gains rather than comprehension improvement. Although not our main point, the inadequacy of our comprehension tests is an important study limitation. Although we could say that we had few alternatives or that many researchers also struggle with inadequate tests of comprehension, such considerations do little to mitigate the seriousness of using inappropriate tests (cf. Cutting & Scarborough, 2006; Keenan et al., 2008).

Moderation analyses.—Whereas our first study aim was to explore the Nashville Early Reading Program's efficacy and whether a COMP component added value to DF, a second purpose was to determine if the program affected first-grade participants *uniformly* or *differentially* such that one subgroup of the sample benefitted more than another. We found evidence of the latter, although inconsistent evidence, perhaps because of relatively low power to detect such relations. Generally, children in DF and DF+COMP with weaker word reading outperformed similarly low-performing controls to a greater extent than treatment children with stronger word reading outperformed comparable controls. For this interaction, *p*-values were 0.02, 0.11, and 0.09 for word reading, non-word reading, and reading comprehension, respectively.

We suspect we obtained these findings because lower-skilled readers in DF and DF+COMP groups were in greater need of the tutors' systematic presentation of appropriate content and their frequently expressed encouragement. Stronger readers in DF and DF+COMP, we believe, were better able to profit from classroom instruction.

There was some suggestion of an additional disordinal relation: one between DF versus DF +COMP and pretreatment word reading such that the DF+COMP condition benefitted poorer readers, whereas DF was more helpful to stronger readers. Although we observed this pattern across word reading, non-word reading, and comprehension outcomes, no interaction was significant. Moreover, the nature of these interactions ran contrary to what we expected prior to study implementation.

Aptitude-by-Treatment Interaction (ATI)

Like many program developers, we typically set out to produce instructional regimens with robust effects, effects that benefit virtually all children of a certain age- or grade level and in an educational setting like the general classroom. Use of moderator analysis reflects recognition that, despite instructional researchers' hopes and ambitions, their programs can vary in value for different subgroups. The programs can provide substantial help to some while offering little benefit to others. In this study, we used word reading as a moderator because of its acknowledged role in beginning reading and because many children struggle at the word level. Results suggested that the Nashville Early Reading Program benefited the youngsters in our sample with lower pretreatment word reading, but not those with stronger pretreatment word reading despite that they, too, were at risk.

Experimental and correlational psychology.

Sixty years ago, Lee Cronbach used different terms to explain the importance of this more complicated view of treatment programs and their effects. In his influential 1957 paper in *American Psychologist*, he called for integrating the two traditions of experimental and correlational psychology. “The experimental method,” he explained, “brings situational variables under tight control, permitting rigorous tests of hypotheses and confident statements about causation” (p. 672). “[Its purpose] is to modify treatments so as to obtain the highest performance when all persons are treated alike—a search, that is, for the ‘one best way’ “ (p. 678). By contrast, he continued, “[t]he correlational method can study what man has not learned to control or can never hope to control” (p. 672). “[Its purpose] is to raise average performance by treating persons differently—different job assignments, different therapies, different disciplinary methods. The correlationist [sic] is utterly antagonistic to a doctrine of the one best way” (p. 678), and “is in love with just those variables that the experimenter left at home to forget” (p. 674).

Cronbach declared it was time for a merger of the experimental and correlational traditions. “If kept independent,” he warned, “they can give only wrong answers or no answers at all regarding certain important problems” (p. 673). He proposed that “the manipulating and correlating schools of research…crossbreed, to bring forth a science of [ATI]” (Cronbach, 1975, p. 116). “Applied psychologists,” he said, “should deal with treatments and persons. Treatments are characterized by many dimensions [.] [S]o are persons. The two sets of dimensions together determine a payoff surface. For any practical problem, there is some best group of treatments to use and some best allocation of persons to treatments. Ultimately, we should *design* treatments, not to fit the average person, but to fit groups of students with particular aptitude patterns” (Cronbach, 1957, 680–681).

So, piggybacking on Cronbach's colloquial language, moderation analysis helps program developers identify the variables most experimenters left at home to forget (p. 674), which the developers may then crossbreed with treatments to find a program most efficacious for a select group of children. Importantly, moderator analysis yields correlational results. Finding that x moderated y requires corroboration through formal experimentation. Moderation analysis is the beginning of a line of work, not its endpoint. It isn't a synonym of ATI. Hence, our finding that the Nashville Early Reading Program appeared to exert differential effects among a young, at-risk sample may not be taken at face value. Its real value awaits experimentation, complete with hypotheses and *a priori* contrasts that involve specified treatments and learners.

Past and future.

Decades ago, special education researchers constructed a class of treatments that targeted specific cognitive deficits among low-achieving special-needs students. The treatments, they promised, would remediate the children's cognitive deficits and strengthen their academic performance. A well-known product from this endeavor was the Illinois Test of Psycholinguistic Abilities (ITPA; e.g., Kirk, McCarthy, & Kirk, 1968). Its developers depended on an information processing model as inspiration and guide to generate learner deficits in areas like "visual decoding" and "auditory associations." Educators were encouraged to use the ITPA to assess children's strengths and weaknesses in these areas and, once a deficit was identified, to find a corresponding treatment within its pages to strengthen it. Different deficits called for different treatments. By the early 1970s, the ITPA was one of many programs claiming the capacity to help instructors match students' cognitive deficits with treatments designed to remediate them. These programs became popular nationwide and they collectively represented the "abilities training" movement.

A key assumption of the developers of these programs was that children's cognitive deficit(s) had to be strengthened prior to engaging them in academic instruction, which often resulted in long delays before they received the intensive direct instruction in reading, math, writing, and so forth that they needed. Sadly, the abilities training movement, despite its popularity, was a profoundly naïve and flawed attempt at ATI that hurt rather than helped many academically-vulnerable students. By the end of the 1970s, it had been convincingly debunked (cf. Mann, 1979). Its assessments of cognitive deficits were shown to lack validity; its treatments to strengthen the putative deficits and to improve academic performance were proved ineffectual (e.g., Arter & Jenkins, 1979).

With the repudiation of the abilities training movement, many also rejected ATI (cf. Reschly & Ysseldyke, 1995). Ever since, the two have often been linked, understood as a joint phenomenon, described as a cautionary tale for those who would take account of individual differences when developing instructional programs. This despite that researchers are (a) finding ATIs in various areas of inquiry (cf. L. Fuchs et al., 2014), and (b) recognizing the very real limits of direct instruction as they work to help students with serious learning problems.

The longstanding view of ATI as an intellectually bankrupt concept, we believe, is tantamount to throwing the baby out with the bath. In comparison to the developers of abilities training programs decades ago, researchers today understand that ATI research is a difficult multi-step process requiring substantive expertise, technical sophistication, humility, and persistence. ATI will be an important dimension of the future of instructional psychology and educational practice, despite what many educational researchers may think. Medical researchers have taken hold of the basics of ATI with both hands. They call it "personalized medicine."

Appendix A: Measures

Screening Measures

Alphabetic knowledge.

The Rapid Letter Naming Test (Fuchs, D., Fuchs, L.S., Thompson, A., Al Otaiba, S., Yen, L., Yang, N.J., Braun, M., & O'Connor, 2001) measures the number of letters named correctly in 60 seconds. All capital and lowercase letters of the alphabet are presented in random order. *The Rapid Sound Naming Test* (D. Fuchs et al., 2001) measures the number of letter sounds named correctly in 60 seconds. The 26 letters of the alphabet are presented in random order. If there is no response in 3 seconds on the Letter Naming and Sound Naming Tests, the tester names the letter or sound and the student moves to the next item. Scores are adjusted if students finish in less than 60 seconds. Test-retest reliability for the Rapid Sound Naming Test is .92 among first-grade students (L. Fuchs & Fuchs, 2001).

Word reading.

The Sight Word Efficiency subtest of the TOWRE (Torgesen, Wagner, & Rashotte, 1999) measures the number of sight words identified correctly in 45 seconds. It consists of 104 sight words arranged from easiest to most difficult. The manual reports an alternate-form coefficient of .97 for a sample of 6 year olds. *Word Identification Fluency A (*WIF A; L. Fuchs, Fuchs, & Compton, 2004) comprises 2, 50-word lists. Words on each list were randomly selected from 133 high-frequency words of the Dolch pre-primer, primer, and first-grade-level lists. The 50 words were printed randomly on a page; students have 60 sec to read each list; and their score is the mean number of words read correctly across them. If a student requires less than 60 sec, the score is pro-rated. The range of scores in fall of first grade in a representative local sample was 0 to 67, for which the mean was 27 (*SD* = 15). Alternate-form reliability is .95 to .97 at first grade (L. Fuchs et al., 2004). *The Word Identification subtest of the WRMT* (Woodcock, 1998) consists of 106 sight words, arranged from easiest to most difficult. Testing is discontinued when the student gives six consecutive incorrect answers. The manual reports the split-half reliability coefficient for a first-grade sample as .98. Cronbach's alpha computed on the sample's pre-treatment performance was . 95.

Non-word reading.

The Phonemic Decoding Efficiency subtest of the TOWRE (Torgesen et al., 1999) includes 63 pseudo-words (e.g., *pim*) presented from easiest to most difficult. The test determines the number of items a student identifies correctly in 45 seconds. Cronbach's alpha based on the sample's post-treatment performance was .91. Alternate-form reliability, reported in the manual for a sample of 6-years-olds, is .97. *The Word Attack subtest of the WMRT* (Woodcock, 1998) consists of 45 pseudo-words, arranged from easiest to most difficult. Testing is discontinued when the student incorrectly answers 6 consecutive items. The manual reports the split-half reliability for a first-grade sample as .94.

IQ.

IQ was assessed by the *Vocabulary and Matrix Reasoning subtests of the WASI* (Wechsler, 1999). Vocabulary consists of 37 items that assess expressive vocabulary, verbal knowledge, memory, learning ability, and crystallized and general intelligence. Participants identify pictures and define words. Cronbach's alpha for the sample's pre-treatment performance was .83. Test-retest reliability, as reported in the manual for 6 to 11 year olds, is .85. Matrix Reasoning includes 32 items that measure nonverbal fluid reasoning and general intelligence. Participants select 1 of 5 options that best completes a visual pattern. Cronbach's alpha for the sample at pre-treatment testing was .84. The manual states that test-retest reliability for children between the ages of 6 and 11 is .76.

Outcome Measures

Outcome measures included the above-mentioned word-reading and non-word reading tests (except WIF A) as well as the following measures. *Word Identification Fluency B* (WIF B; Zumeta, Compton, & Fuchs, 2012)) word lists were drawn from the 500 most frequently written words compiled by Zeno et al. (1995). Zumeta et al. divided the 500 words into 50 groups of 10 such that group #1 consisted of most frequently appearing words; group #2, the next most frequently used words; the last group consisted of least frequently written words. Zumeta et al. then repeatedly and randomly selected 1 word from each of the 10-word groups to create 20 lists of 50 words, which were positioned on a page with most common words first. Alternate form reliability was .95 to .97. In this study, two alternate forms of WIF B were administered to each child in fixed order for 18 weeks. Week #1 scores were "pretreatment"; Week #18 scores were "posttreatment." Administration and scoring were the same as WIF A.

The Passage Comprehension subtest of the WRMT (Woodcock, 1998) is a norm-referenced, modified cloze procedure. For the first set of items, the tester presents a rebus, and asks the child to point to the picture corresponding to it. Next, the child points to the picture representing words printed on the page. Later items require silently reading a passage and identifying a missing word. The test is discontinued after 6 consecutive incorrectly answered items. Split-half reliability at first grade is .94 (Woodcock, 1998). Cronbach's alpha based on the sample's posttreatment performance was .79. *The Reading Comprehension subtest of the ITBS* (Hoover, Dunbar, & Frisbie, 2001) has 3 sections: Sentences, Picture Story, and Story. For Sentences, students read a sentence with a missing word and select 1 of 3 words to complete it. Picture Story requires them to answer 4 questions about a picture. For each question, they select 1 of 3 answers. Story section items require students to read a short passage followed by 4 questions. They respond to each by selecting an answer from 3 options. The test consists of 19 items. Cronbach's alpha calculated on the sample's posttreatment performance was .82.

Additional Measures

The *DF placement test* helped tutors determine the book level most appropriate to begin instruction for each student. The test consists of 125 items, divided into five increasingly difficult bands of 25 items. Each band is aligned with a book level. The placement test has two item types: grapheme-phoneme correspondences (GPCs) and decodable words. Students were tested on every GPC presented in tutoring. The number presented (and tested) ranges from 20 in Book 1 to 1–5 in subsequent books. Decodable words are selected on the basis of how representative they are of all words in a book. “Representativeness” is defined as how frequently a word is used and how many letters are in the word. If a student reaches 6 incorrect items (in any band), the examiner stops the test and records the number of items missed. If students commits 5 or fewer errors, they move to the next higher (more difficult) band.

This research was supported in part by grant #R01HD056109 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD). The authors are responsible for the paper’s content, which does not necessarily represent the views of the NICHD or the National Institutes of Health. We thank Brett Miller (NICHD) for his important advice and encouragement; Rebecca Zumeta Edmonds for her normative data on Word Identification Fluency B; and the teachers, principals, and administrators of the Metro Nashville Public Schools for their interest and cooperation.

References

Aguinis H (1995). Statistical power problems with moderated multiple regression in management research. Journal of Management, 21, 1141–1158.

Arter JA, & Jenkins JR (1979). Differential diagnosis—prescriptive teaching: A critical appraisal. Review of Educational Research, 49, 517–555.

Baker L, & Brown AL (1984). Metacognitive skills and reading In Pearson PD (Ed.), Handbook of reading research, pp.394–535. New York: Plenum Press.

Brown R, Pressley M, Van Meter P, & Schuder T (1996). A quasi-experimental validation of transactional strategies instruction with low-achieving second-grade readers. Journal of Educational Psychology, 88(1), 18.

Case LP, Speece DL, Silverman RD Schatschneider C, Montanaro E, & Ritchey KD (2014). Immediate and long-term effects of Tier 2 reading instruction for first grade students with a high probability of reading failure. Journal of Research on Educational Effectiveness, 7*(1),* 28–53.

Catts HW, Fey ME, Tomblin JB, & Zhang X (2002). A longitudinal investigation of reading outcomes in children with language impairments. Journal of Speech, Language, and Hearing Research, 45(6), 1142–1157.

Cohen J, Cohen P, West SG, & Aiken LS (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Connor CM, Morrison FJ, & Katch LE (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. Scientific Studies of Reading, 8(4), 305–336.

Cronbach LJ (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671–684.

Cronbach LJ (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116–127.

Cutting LE, & Scarborough HS (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. Scientific Studies of Reading, 10(3), 277–299.

Elleman AM (2017). Examining the Impact of Inference Instruction on the Literal and Inferential Comprehension of Skilled and Less Skilled Readers: A Meta-Analytic Review. Journal of Educational Psychology.

Elleman AM, Lindo EJ, Morphy P, & Compton DL (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. Journal of Research on Educational Effectiveness, 2, 1–44.

Flavell JH (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American Psychologist, 34(10), 906.

Foorman BR, Francis DJ, Shaywitz SE, Shaywitz BA, & Fletcher JM (1997a). The case for early reading interventions In Blachman BA (Ed.), Foundations of reading acquisition and dyslexia: Implications for early intervention (pp. 243–264). Hillsdale, NJ: Erlbaum.

Fuchs D, Fuchs LS, Elleman AM, Peng P, Kearns DM, Gilbert K, Compton DL, Patton S, Steacy LM, Toste JR, & Miller AC (2018). A randomized control trial of explicit instruction with and without cognitive training to strengthen the reading comprehension of at-risk first-grade children. Submitted.

Fuchs D, Fuchs LS, Mathes PG, & Simmons DC (1997). Peer-Assisted Learning Strategies: Making classrooms more responsive to diversity. American Educational Research Journal, 34, 174–206.

Fuchs D, Fuchs LS, Thompson A, Al Otaiba S, Yen L, Yang NJ, Braun M, & O' Connor RE (2001). Is reading important in reading-readiness programs: A randomized field trial with teachers as program implementers. Journal of Educational Psychology, 93(2), 251–267.

Fuchs D, Kearns D, Elleman A, Fuchs LS, et al. (2015). The Nashville early reading project: A manual. Nashville: Vanderbilt University.

Fuchs LS, & Fuchs D (2001). Progress monitoring with letter sound fluency: Technical data. Available from L.S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37220.

Fuchs LS, Fuchs D, & Compton DL (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. Exceptional Children, 71, 7–21.

Fuchs LS, Schumacher RF, Sterba SK, Long J, Namkung J, Malone A, Hamlett CL, Jordan NC, Gersten R, Siegler RS, & Changas P (2014). Does working memory moderate the effects of fraction intervention? An aptitude-treatment interaction. Journal of Educational Psychology, 106, 499–514. doi: 10.1037/a0034341)

Gersten R, Newman-Gonchar RA, Haymond KS, & Dimino J (2017*). What is the evidence base to support reading interventions for improving student outcomes in grades 1–3?* (REL 2017–271). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast Retrieved from http://ies.ed.gov/ncee/edlabs.

Hoover HD, Dunbar SB, & Frisbie DA (2001). Iowa Tests of Basic Skills. Itasca, IL: Riverside.

Keenan J, Betjeman R, & Olson R (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. Scientific Studies of Reading, 12, 281–300.

Kendeou P, Broek P, Helder A, & Karlsson J (2014). A cognitive view of reading comprehension: Implications for reading difficulties. Learning Disabilities Research & Practice, 29(1), 10–16.

Kirk SA, McCarthy JJ, & Kirk WD (1968). Illinois Test of Psycholinguistic Abilities. Urbana: University of Illinois Press

Lang LI, Torgesen J, Vogel W, Chante C, Lefsky E, & Petscher Y (2009). Exploring the relative effectiveness of reading interventions for high school students. Journal of Research on Educational Effectiveness, 2, 149–175.

Mann L (1979). On the trail of process: A historical perspective on cognitive processes and their training. New York: Grune & Stratton.

McGee A, & Johnson H (2003). The effect of inference training on skilled and less skilled comprehenders. Educational Psychology, 23(1), 49–59.

McNeish DM, & Stapleton LM (2016). The effect of small sample size on two-level model estimates: A review and illustration. Education Psychology Review, 28, 295–314. doi: 10.1007/s10648-014-9287-x

Miller B, McCardle P, & Hernandez R (2010). Advances and remaining challenges in adult literacy research. Journal *of Learning Disabilities*, 43(2), 101–107. [PubMed: 20179305]

Morrow LM, Rand MK, & Young J (1997*).* Differences between social and literacy behaviors of first, second, and third graders in social cooperative literacy settings. New Brunswick, NJ: Rutgers University.

Palinscar AS, & Brown AL (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction, 1(2), 117–175.

Perfetti C (2007). Reading ability: Lexical quality to comprehension. Scientific Studies of Reading, 11(4), 357–383.

Perfetti CA, & Hart L (2001). The lexical bases of comprehension skill In Gorfien DS (Ed.), On the consequences of meaning selection: Perspectives on resolving lexical ambiguity (pp. 67–86). Washington, DC: American Psychological Association.

Perfetti CA, Marron MA, & Foltz PW (1996). Sources of comprehension failure: Theoretical perspectives and case studies In Cornoldi C & Oakhill J (Eds.), Reading comprehension difficulties: Processes and intervention (pp. 137–165). Mahwah, NJ: Erlbaum.

Raphael TE (1984). Teaching learners about sources of information for answering comprehension questions. Journal of Reading, 27(4), 303–311.

Reschly DJ, & Ysseldyke JE (1995). School psychology paradigm shift (pp. 17–31). In Thomas A & Grimes J (Eds.), Best practices in school psychology. Washington, DC: National Association of School Psychology.

Roberts C, & Roberts SA (2005). Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials, 2, 152–162. doi: 10.1191/1740774505cn076oa [PubMed: 16279137]

Scott CM (2009). A case for the sentence in reading comprehension. Language, Speech, and Hearing Services in Schools, 40(2), 184–191.

Shanahan T, Callison K, Carriere C, Duke NK, Pearson PD, Schatschneider C, & Torgesen J (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010–4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education Retrieved from whatworks.ed.gov/publications/practiceguides.

Schunemann N, Sporer N, & Brunstein JC (2013). Integrating self-regulation in whole-class reciprocal teaching: A moderator-mediator analysis of incremental effects on fifth graders' reading comprehension. Contemporary Educational Psychology, 38, 289–305.

Stram DO, & Lee JW (1994). Variance components testing in the longitudinal mixed effects model. Biometrics, 50, 1171–1177. Retrieved from http://www.jstor.org/stable/2533455. [PubMed: 7786999]

Torgesen JK (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. Learning Disabilities Research & Practice, 15(1), 55–64.

Torgesen JK, Wagner RK, & Rashotte CA (1999). Test of Word Reading Efficiency. Austin, TX: PRO-ED.

Torgesen JK, Wagner RK, Rashotte CA, Rose E, Lindamood P, Conway T, & Garvan C (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. Journal of Educational Psychology, 91(4), 579.

Vadasy PF, & Sanders EA (2011). Efficacy of supplemental phonics-based instruction for low-skilled first graders: How language minority status and pretest characteristics moderate treatment response. Scientific Studies of Reading, 15(6), 471–497.

Wanzek J, Wexler J, Vaughn S, & Ciullo S (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. Reading and Writing, 23(8), 889–912. [PubMed: 21072128]

Wechsler D (1999). Wechsler Abbreviated Scale of Intelligence. San Antonio, TX: The Psychological Corporation.

Woodcock RW (1997). Woodcock Diagnostic Reading Battery. Itasca, IL: Riverside.

Zeno SM, Ivens SH, Millard RT, & Duvvuri R (1995). The educator's word frequency guide. Brewster, NY: Touchstone Applied Science.

Zumeta RO, Compton DL, & Fuchs LS (2012). Using Word Identification Fluency to monitor first-grade reading development. Exceptional Children, 78(2), 201–220. [PubMed: 22736804]

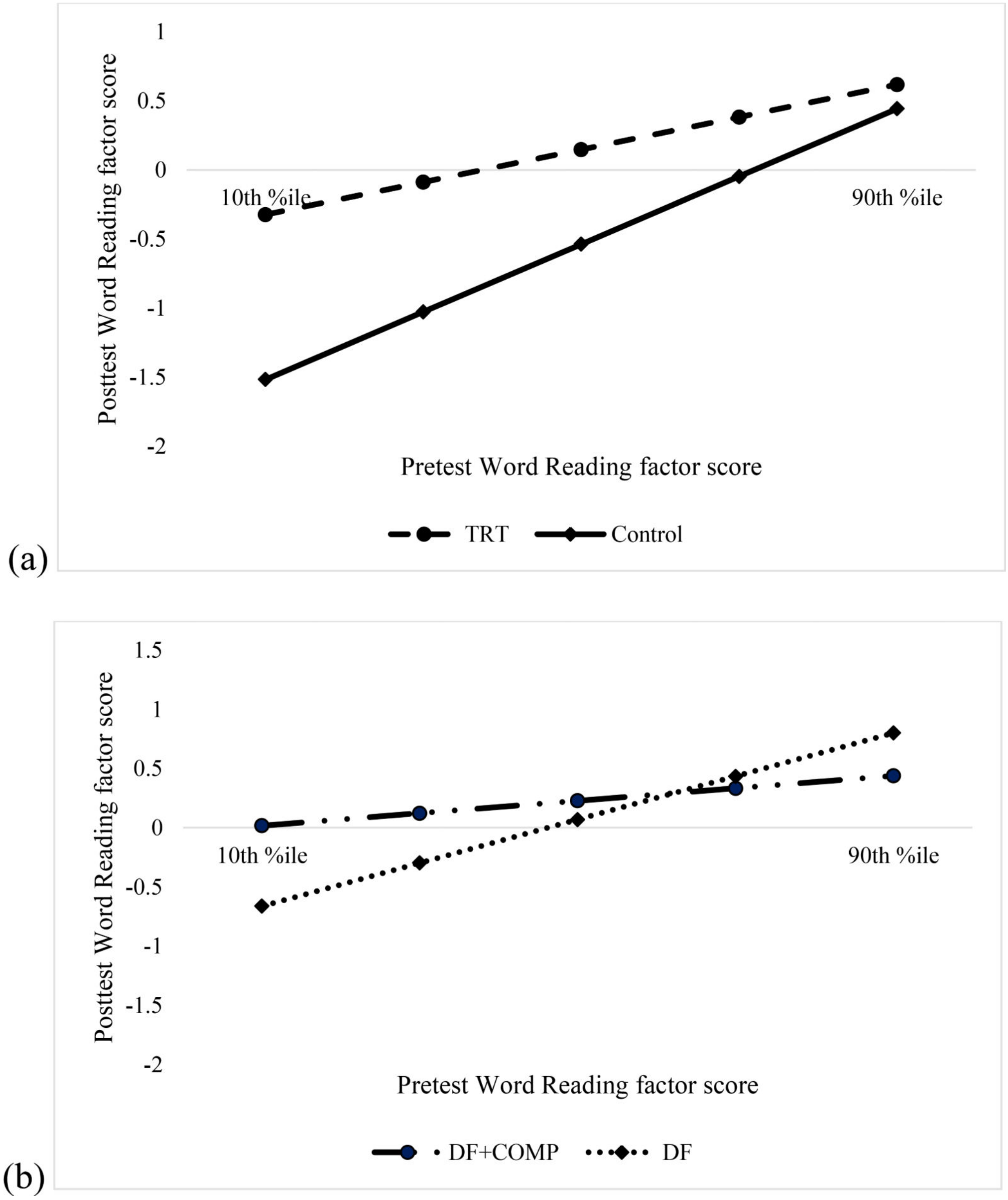

Figure 1.

Interaction graphs from multilevel model for the Word Reading outcome examining pretreatment Word Reading as a moderator. The first graph (1a) represents the moderated treatment vs. control effect and the second (1b) represents the moderated effect of DF +COMP vs. DF.

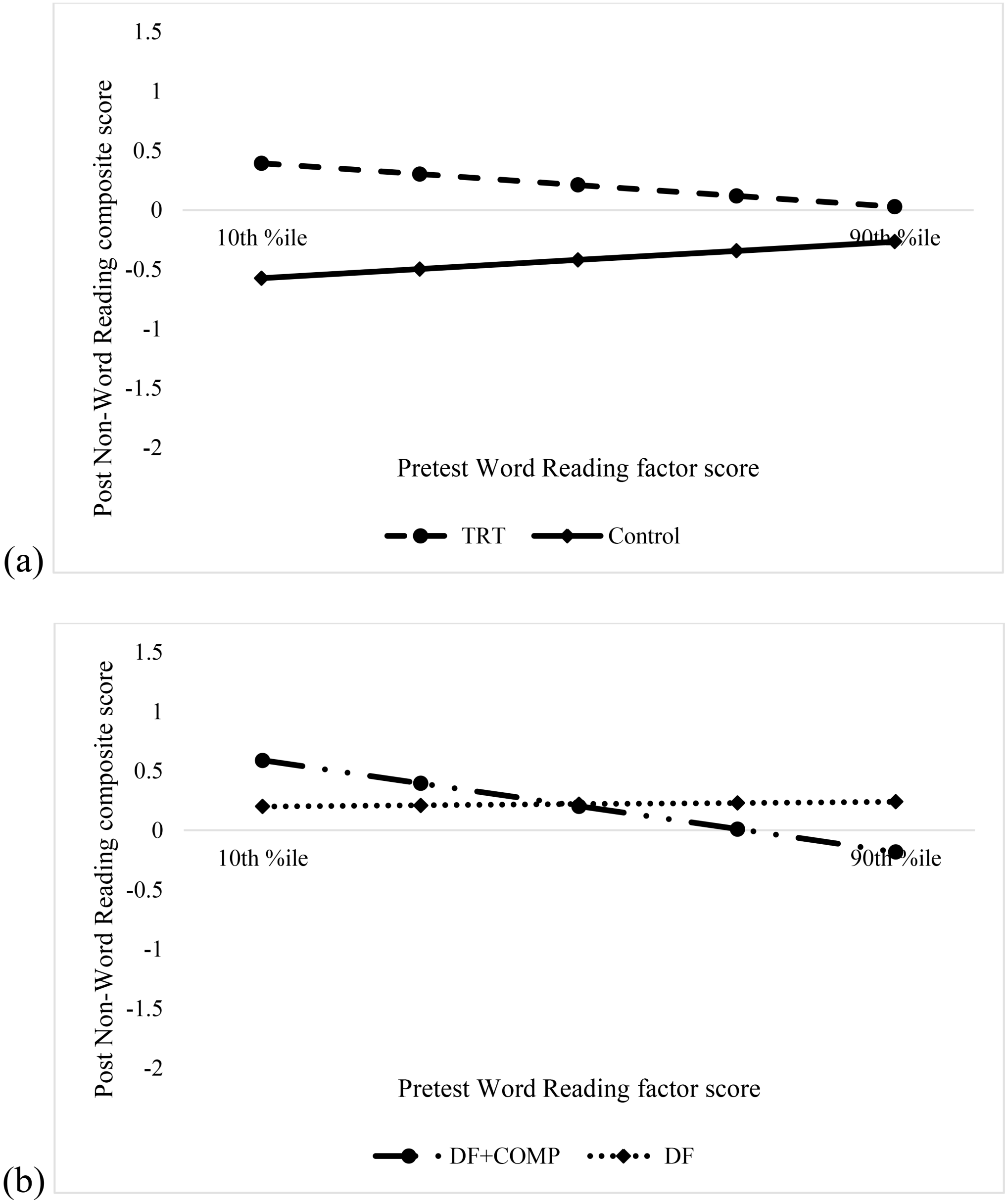

Figure 2.

Interaction graphs from multilevel model for the Non-Word Reading outcome examining pretreatment Word Reading as a moderator. The first graph (2a) represents the moderated treatment vs. control effect and the second (2b) represents the moderated effect of DF +COMP vs. DF.

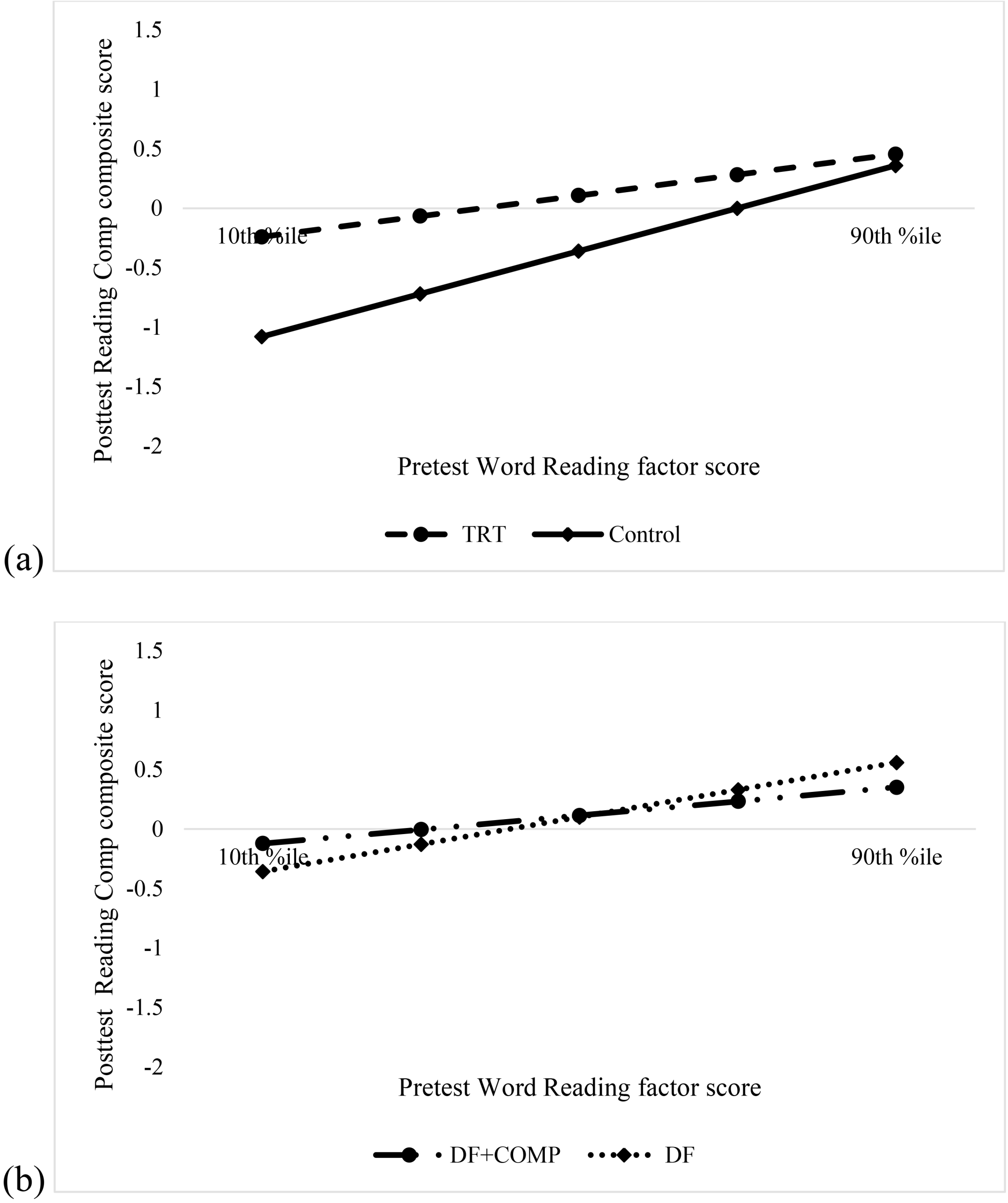

Figure 3.

Interaction graphs from multilevel model for the Reading Comprehension outcome examining pretreatment Word Reading as a moderator. The first graph (3a) represents the moderated treatment vs. control effect and the second (3b) represents the moderated effect of DF+COMP vs. DF.

Table 1

Student Demographics by Study Condition

	DF (*n* = 43)		DF+COMP (*n* = 40)		Control (*n* = 42)	
	f	**%**	***F***	**%**	***f***	**%**
Males	24	57.14	27	69.23	21	50.00
Student Race						
Black	13	30.95	16	41.03	18	42.86
White	14	33.33	11	28.21	10	23.81
Hispanic	8	19.05	8	20.51	8	19.05
Other	7	16.67	4	10.26	6	14.29
Free/Reduced Lunch	36	85.71	32	82.05	38	90.48
IEP Students	1	2.78	2	6.06	3	7.89
ELL Students	6	14.63	7	18.92	8	19.05
Retained Students	5	12.20	1	2.56	3	7.14

Note. Percentages are based on the number of students in each condition with reported demographic data.

Table 2

Pre- and Posttreatment Means (Standard Deviations) and Percentile Scores on Reading Measures by Study Condition

	DF (*n* = 43)				DF+COMP (*n* = 40)				Control (*n* = 42)			
Domain/Measure	**Pre**	**%tile**	**Post**	**%tile**	**Pre**	**%tile**	**Post**	**%tile**	**Pre**	**%tile**	**Post**	**%tile**
Word Reading												
TOWRE SWE	10.95 (4.95)	47.5	34.02 (11.12)	65.9	9.93 (3.96)	45.4	34.28 (10.16)	66.4	9.57 (4.10)	44.4	27.36 (11.03)	54.9
WRMT WID	11.28 (5.81)	59.4	38.33 (9.53)	64.2	9.73 (5.43)	55.7	38.58 (7.64)	65.1	9.76 (5.58)	55.4	30.76 (9.94)	50.8
WIF A	6.08 (3.02)	6.9	25.64 (11.57)	–	5.67 (2.40)	5.9	24.80 (9.70)	–	5.71 (3.06)	5.9	16.59 (9.35)	–
WIF B												
Non-Word Reading												
WRMT WA	2.58 (2.95)	46.1	13.72 (6.94)	61.9	2.53 (2.81)	47.7	13.25 (7.48)	62.1	2.29 (2.65)	44.3	8.00 (5.72)	42.3
TOWRE PDE	2.42 (2.86)	48.0	10.67 (7.21)	49.8	2.78 (3.03)	50.2	10.68 (6.29)	51.4	2.48 (2.99)	48.7	6.74 (4.99)	37.8
Reading Comprehension												
WRMT-PC	9.12 (2.50)	41.3	16.33 (3.06)	46.8	8.88 (3.01)	39.6	15.70 (2.99)	43.2	8.33 (2.66)	36.2	14.10 (3.18)	35.9
ITBS-RC	5.84 (2.85)	24.9	13.53 (4.16)	43.6	6.03 (1.87)	26.7	13.41 (3.48)	42.1	5.71 (2.61)	23.6	11.10 (4.44)	30.3

Note. Passage Comp = Woodcock Reading Mastery Test-Revised Passage Comprehension; ITBS = Iowa Test of Basic Skills Reading Comprehension; SWE = Test of Word Reading Efficiency-Sight Word Efficiency; WID = Woodcock Reading Mastery Test-Revised - Word Identification; WIF = Word Identification Fluency; WA = Woodcock Reading Mastery Test-Revised - Word Attack; PDE = Test of Word Reading Efficiency-Phonemic Decoding Efficiency.

Table 3

Effects of Treatment and Pre-treatment Word Reading across Reading Outcomes (N = 124)[a]

	Word Reading			Non-Word Reading			Reading Comprehension		
	Est.	***SE***	***p***	**Est.**	***SE***	***p***	**Est.**	***SE***	***p***
Fixed Effects									
Intercept, γ_0	−0.02	0.09	.800	−0.01	0.08	.940	−0.01	0.08	.938
Pretreatment score, γ_1	—	—	—	0.39	0.10	<.001	0.31	0.16	.057
Treatment vs. Control, γ_2	0.64	0.15	<.001	0.60	0.14	<.001	0.43	0.13	.002
DF+COMP vs. DF, γ_3	0.11	0.18	.526	−0.05	0.20	.784	−0.01	0.16	.968
Pre-treatment Word Reading, γ_4	0.47	0.08	<.001	−0.05	0.09	.573	0.34	0.08	<.001
Treatment vs. Control × Pre-treatment Word Reading, γ_5	−0.37	0.16	.020	−0.26	0.16	.111	−0.28	0.16	.086
DFCOMP vs. DF × Pre-treatment Word Reading, γ_6	−0.36	0.19	.055	−0.31	0.23	.186	−0.17	0.20	.371
Random Effects									
School, var(u_{0k})	0.02			0.00			0.03		
Classroom, var(r_{0jk})	0.04			0.03			0.00		
Residual, var(e_{ijk})	0.54			—			0.46		
Residual, var(e_{ijk}|DF)	—			0.68			—		
Residual, var(e_{ijk}|DFCOMP)	—			0.76			—		
Residual, var(e_{ijk}|Control)	—			0.34			—		

[a] *N* = 124 due to missing test score.