Evaluating a Multidimensional Reading Comprehension Program and Reconsidering the Lowly Reputation of Tests of Near Transfer

Douglas Fuchs, Emma Hendricks, Meagan E. Walsh, Lynn S. Fuchs, Jennifer K. Gilbert,

Wen Zhang Tracy, Samuel Patton III, Nicole Davis, & Wooliya Kim

Peabody College of Vanderbilt University

Amy M. Elleman

Middle Tennessee State University

Peng Peng

University of Nebraska-Lincoln

Final report submitted to publisher 11/21/2017

In press

Learning Disabilities Research & Practice 2018

This research was supported in part by a grant (R23413D0003) from the National Center for Special Education Research in the Institute of Education Sciences (U.S. Department of Education) and Core Grant HD 15052 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. This paper does not necessarily reflect positions or policies of the funding agencies, and no official endorsement by them should be inferred.

Abstract

We conducted a 14-week experimental study of 2 versions of a relatively comprehensive RC intervention that involved 50 classroom teachers, 15 tutors, and 120 children drawn in equal proportions from grades 3 and 5 in 13 schools in a large urban school district. Students were randomly assigned in equal numbers to the 2 tutoring conditions and a control group. Results indicated that students in the 2 tutored groups tended to perform comparably on all tests and to outperform controls (more so in grade 5 than grade 3) on near-transfer but not far-transfer measures of RC. This differential pattern of program effects for near- versus far-transfer measures raises questions about how tests of near-transfer and far-transfer are conventionally understood.

Evaluating a Multidimensional Reading Comprehension Program and Reconsidering the Lowly Reputation of Tests of Near Transfer

Many students in the United States do not read with adequate understanding. In 2015, only 36% of 4th-grade students scored "at or above proficient" (i.e., grade level or better) on the National Assessment of Educational Progress (NAEP; U.S. Department of Education 2016), a percentage that hasn't changed appreciably since 1992. Children with inadequate reading comprehension (RC) struggle to understand both narrative and informational texts, and their difficulty in understanding informational texts ensures poor performance in content classes throughout their school career (Meneghetti et al., 2006). Hence, there is a need for effective RC instruction, especially that which aims to strengthen students' comprehension of informational texts.

Explicit instruction has long been the gold standard for helping at-risk children learn to read (Becker, Englemann, Carnine, & Rhine, 1981). Whereas a large body of research indicates that it improves word-reading (National Reading Panel, 2000), its impact on RC is less clear. This is probably because RC is a multidimensional construct (Kintsch & Kintsch, 2005; Nation, 2005), requiring word recognition, reading fluency, vocabulary and background knowledge, working memory, verbal reasoning, sensitivity to language structures, strategy use (e.g., retelling, constructing main ideas, and self-monitoring), and more. Researchers and program developers are uncertain as to which skills, cognitive processes, and strategies should be included in a given RC intervention, and how they should be combined into a cohesive whole that practitioners would find feasible to use on a regular basis (Fuchs et al., 2016).

Notwithstanding such uncertainty, RC researchers have often focused on strategy instruction to help students build "situation models." As an example, children may be taught to activate their background knowledge (e.g., Compton, Miller, Gilbert, & Steacy, 2013) and combine it with inferential reasoning (Kintsch, 1988) to continuously update known information with new information. Strategy instruction has been shown to improve the RC of many elementary-age students (Gersten, Fuchs, Williams, & Baker, 2001).

Nevertheless, a sizable minority of children do not benefit from such instruction (O'Connor & Fuchs, 2013). These students require something more or different. For the last several years, we have been funded as a Research Initiative by the National Center for Special Education Research in the Institute of Education Sciences (U.S. Department of Education) to (a) develop a relatively comprehensive instructional program that teaches RC strategies and (b) embed cognitive training in the "base" program to strengthen processes strongly linked to RC.

One cognitive process on which we have focused is working memory (WM). Studies show a positive relation between it and RC (Berninger & Richards, 2010; Carretti, Caldarola, Tencati, & Cornoldi, 2014; Daneman & Merikle, 1996; Palladino et al., 2001). Moreover, poor readers, including many with learning disabilities, often demonstrate weak performance on both WM and RC (Swanson & Howell, 2001; Swanson & O'Connor, 2009). We hypothesize that enhancing WM capacity should strengthen RC beyond improvements expected from an RC intervention that incorporates a comprehensive set of strategies.

Many have attempted to strengthen WM with a variety of domain-general tasks (cf. Schwaighofer, Fischer, & Buhner, 2015) in hopes that a fortified WM would transfer to, or facilitate, improved performance in multiple academic domains, including RC. Whereas these training efforts have sometimes improved WM performance, they have not improved reading outcomes (Melby- Lervåg & Hulme, 2013; Schwaighofer et al., 2015). In contrast to previous efforts, we have been pursuing a *domain-specific* approach to WM training to increase the likelihood of transfer to RC performance. A domain-specific approach reflects a belief that WM is best understood as closely related to skills and knowledge specific to a given domain (Unsworth & Engle, 2007). Accordingly, WM should be trained as part of domain-specific activities. We have been developing a WM training program and embedding it in the context of RC tasks. Findings from at least two prior studies suggest that a domain-specific approach to WM training may be beneficial (Garcia-Madruga et al., 2013; Carretti et al., 2014).

The purpose of our study was two-fold. First, we explored the efficacy of a multidimensional tutoring program to strengthen RC in informational texts. Second, we looked at effects of a variation of this program that included a WM training component. Accordingly, we created two active treatment conditions—the RC program alone (hereafter, "COMP") and COMP combined with WM training ("[WM]COMP")—and a control condition. We investigated the efficacy of the treatments against controls in a 14-week randomized trial with third- and fifth-grade at-risk learners during the 2015-2016 academic year.

Method

Participants

Teachers. Students meeting the criteria for study inclusion (see below) came from 50 classrooms in 13 schools in a large school district in the Southeast. Table 1 shows demographic data for the 50 teachers. Most were Caucasian and female, and a majority held a graduate degree. Half of the third-grade teachers were certified to teach English Language Learners, one-fifth were certified in reading, and fewer were certified to teach special education. At grade 5, half of the teachers were certified in reading; one-fifth were certified to teach English Language Learners.

Student recruitment and selection. Teachers nominated 500 students who they believed (a) were at risk for reading difficulties, (b) did not have severe behavior problems or developmental/intellectual disabilities, and (c) were unlikely to be frequently absent from school. English Language Learners proficient in English (e.g., scoring between 38 and 46 on the Tennessee English Language Assessment) were included in the study. Parents or guardians of 314 of the nominated 500 students gave written permission for their children to participate. Of this group, 304 were screened. Two of 10 not screened refused testing; two more were excluded because of low English language proficiency (despite acceptable test scores); and six were eliminated on the basis of chronically late school arrivals.

A gated screening process facilitated the identification of students scoring in the average range on word reading and below average on RC. We searched for students with adequate word reading skills because we intended our reading intervention to emphasize RC skills and strategies to the maximum extent possible. We did not design our program to strengthen word recognition. The students were first tested on the Sight Word Efficiency (SWE) subtest of the Test of Word Reading Efficiency-2 (TOWRE-2; Torgesen, Wagner, & Rashotte, 2012). Our criteria for children in grades 3 and 5, respectively, were percentile rankings of > 30 and > 13. Those meeting these criteria completed the Reading Comprehension subtest of the Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001) and the Vocabulary and Matrix Reasoning subtests of the Wechsler Abbreviated Scale of Intelligence-2 (WASI-2; Wechsler, 2011). Study participants were required to have ITBS percentile scores < 60 and T-scores > 37 on either of the WASI-2 Vocabulary or Matrix Reasoning subtests.

Final numbers of student participants were 60 at both grade levels. Third-grade students were from 8 schools and 28 classrooms. Fifth-grade students were from 5 schools and 22 classrooms. Third-grade students' mean percentile word reading score on the SWE subtest of the TOWRE was 101.10 (*SD* = 6.23); the comparable score for fifth graders was 94.00 (*SD* = 8.53). On the ITBS, third graders' mean developmental standard score was 166.19 (*SD* = 10.42), which approximates the average score for second-grade children in the test's standardization population. The corresponding averaged ITBS score for fifth graders was 184.88 (*SD* = 14.73), which is the average score of third graders in the ITBS standardization population (see Table 2). Our cut-scores on the TOWRE and ITBS were both higher than we would have liked. But they reflect the difficulty we experienced in identifying students with relatively strong word recognition and weak RC performance.

Assignment to study group. Because the tutoring program was designed to be delivered to pairs of students, it was necessary that both members of a pair could read from the same tutoring materials. To ensure this, and to increase the likelihood that the pretreatment reading scores of students in COMP, [WM]COMP, and control groups were comparable, we used the following procedure to assign students to pairs and then pairs to study groups. First, student pairs were drawn from the same grade in the same school and were matched on SWE scores from the TOWRE. Second, an averaged ITBS standard score was computed for each pair, and the pairs were then placed in strata on the basis of these scores. Third, within strata, block random assignment was conducted to assign the pairs to the three study groups such that 80 students (40 at both grade levels) received tutoring and 40 students (20 at each grade level) were controls.

No differences between study groups (within grade) were obtained on the TOWRE's SWE raw scores or ITBS Reading Comprehension Developmental Standard Scores. Two students from [WM]COMP and two controls moved from the school district prior to post-treatment testing (attrition rate = 3.33%). Because these were the only students who did not complete the study, equivalence statistics comparing “finishers” and “non-finishers” were deemed unnecessary.

Final sample. Table 2 provides demographic and screening data for students completing the study (*n* = 116). Gender was equally represented in grades 3 and 5, and most students in the two grades received free or subsidized lunch. Children of Hispanic ancestry predominated at third grade; African-American children at fifth grade. As mentioned, students in grades 3 and 5 were on average performing one and two grade levels behind, respectively, in RC. Their word reading and cognitive functioning were in the average range.

Project staff. Thirteen research assistants (RAs) worked 20 hours per week. Their project-related responsibilities were to tutor and collect pre- and post-treatment data. Eleven were in master’s programs across several academic departments. Two were doctoral students in special education. Most had prior experience working with children. They were supervised by two full-time project coordinators whose duties included tutoring students. Hence, there were 15 tutors.

Base Treatment Program

Tutoring occurred three times per week for 14 weeks (42 sessions) in a quiet location in the students’ schools. The first and second lessons of each week lasted 45 min and students were tutored in pairs. The third lesson lasted 20 min and was delivered one-to-one. All tutoring sessions were scripted to ensure fidelity of treatment implementation and consistency among the tutors with respect to how tutoring was delivered. However, tutors were trained to use the scripted materials in a relatively “natural” manner.

Students in the two treatment groups received direct instruction in evidence-based strategies, which were intended to strengthen understanding of informational texts. They were taught to use one set of strategies prior to reading, and a different set after reading. Each lesson had the same structure, or organization, and was delivered in the same way. Over the course of the 14-week program, tutors gradually transferred control of strategy use to students. As mentioned, the two active treatment groups, COMP and [WM]COMP, had the same base RC instruction. Children in the [WM]COMP group also participated in three additional activities (described below) that were designed to strengthen WM.

Before-reading strategies. Before reading a passage, students discussed text features with their partners, including title, headings, pictures, tables, and charts. They used the text features to predict the topic of the passage. They also identified bolded vocabulary words and looked up their meanings in a glossary. Next, they read a "guiding question" and made a prediction. Guiding questions were true-false statements or multiple-choice questions. Their purpose was to direct attention to important ideas in the passage. Students reviewed the accuracy of their prediction after they completed reading.

In a final before-reading activity, students were encouraged to probe their background knowledge for relevant information. For some topics, it was unlikely that they had such knowledge. In this circumstance, media (e.g., video clips, audio recordings or cartoons) was used to supply it. After completing these before-reading strategies, partners took turns reading the passage, one or two paragraphs at a time.

After-reading strategies. After reading one or two paragraphs, students re-told the important facts in the order they read them. They were encouraged to use their own words and to begin each statement with a "retell word" (e.g., *first*, *next*, *then*, *last*). Then, they constructed the main idea of the paragraph by identifying the most important person or thing, stating the most important information about the person or thing, and combining information in steps one and two to construct a main idea statement.

They were also taught an “In or Out” strategy to answer factual and inferential comprehension questions. They learned that answers to factual questions are explicitly stated in the passage, whereas answers to inference questions are usually found by combining information from the passage with what they already know (i.e., their background knowledge).

In the first step of the In or Out strategy, students identified “key words” in a question. Key words, they were told, are important because they can help the reader better understand the question and get a sense of whether the answer is inside or outside of the passage. Second, they underlined the key words, as in the following: “How old was Jessica Watson when she started her sailing trip around the world?” Third, they looked for these word groups in the text and then read sentences before and after them. Fourth, if they found the answer to the question in the text, they had to “prove it” to their partner or tutor by pointing to its location. If the answer wasn’t found in the text, they knew the question probably required them to make an inference, which led them to the fifth and final step in the In or Out strategy: They “brainstormed” what they knew about the topic by asking themselves, “What do I know about the key words?,” “What do I remember about them?,” and “Do they remind me of anything?”

Supplemental activities to the base program. Students began each lesson with a speeded cloze activity to build fluency and comprehension. They read a paragraph that summarized the text read during the previous lesson. This paragraph had five to seven semantically or grammatically important words removed and replaced with blank spaces. Next to each blank were two word options. Students circled the word they thought best completed the sentence. Afterward, the tutor reviewed the children’s answers. Beginning in the ninth week of tutoring, students played a game of "What if?" It occurred during the third lesson of each week. The tutor read aloud a short topically related story. The student retold the story. The tutor then introduced a plot twist and assisted the student in creating a new ending.

Working Memory Training

Students in the [WM]COMP group were given the same tutoring as the COMP group with three exceptions: The "What if?" game, and the cloze fluency and main idea activities, were modified to strengthen WM. In the "What if?" game, [WM]COMP students retold the story again, but with a changed ending. This modification required them to update their mental model of the story. After completing the cloze fluency activity, they were required to recall in order the words they had selected to fill in the blanks. After deriving a main idea for a paragraph, they were directed to recall all of the previous main ideas in order, including the one they had just created. In other words, the modifications to the cloze fluency and main idea activities resulted in two kinds of complex span tasks, both of which required students to maintain information in their WM while they processed new information.

Measures

Screening measures. As mentioned, several measures were used to select students for the study. The screening measure for RC was the ITBS Reading Comprehension subtest—Level 9 for grade 3, Level 11 for grade 5 (Hoover et al., 2001)**.** It has a multiple-choice response format, and includes narrative and expository passages of varying length and complexity. It was administered in two sessions. Cronbach's alphas for the study sample were .48 and .52 for third- and fifth-graders, respectively. Our screen for word reading was the SWE subtest of the TOWRE-2 (Torgesen et al., 2012). It assesses the accuracy and fluency with which children read lists of real words in 45 s. Sample-specific reliability is not reported because Cronbach's alpha is an inappropriate index for speeded tests. The manual indicates test-retest reliability is .90 for a sample of third-, fourth-, and fifth-grade children.

We also used two subtests from the WASI-2 (Wechsler, 2011): Matrix Reasoning and Vocabulary. Matrix Reasoning assesses nonverbal reasoning with pattern completion, classification, analogy, and serial reasoning tasks. Sample-based Cronbach's alpha was .83. The Vocabulary subtest evaluates expressive vocabulary, verbal knowledge, and foundational information. Sample-based Cronbach's alpha was .76.

Near-transfer (NT) measures. We developed two NT measures—one testing knowledge acquisition; the other, RC. The NT knowledge acquisition measure required students to answer 20 multiple-choice questions about vocabulary, facts, and ideas in our instructional passages. The measure was administered to all study participants in groups of three to five by a tester "blind" to group affiliations. The tester read aloud questions and answer choices, and proceeded one question at a time so that all students could mark an answer in their test booklets. This test was administered at post-treatment. Sample-based Cronbach's alpha was .69.

The NT RC measure had a multiple-choice format, four expository (informational) passages, and 16 items (i.e., four multiple-choice questions per passage). The measure had the same font and format as that of the passages in the COMP and [WM]COMP treatments. Four of the 16 questions required students to summarize what they read at the passage level; four required them to find factual information stated directly in the text; and the remaining eight questions involved inference-making. Whereas none of the passages had been seen previously by the students, their content was drawn from topics that had been explored during tutoring (e.g., civil rights, inventors). The measure was administered at pre- and post-treatment by a tester unaware of which students were from what study group. Sample-based Cronbach's alpha was .71 and .72 for pretreatment and post-treatment performances, respectively.

Far-transfer (FT) measures. The FT measures of RC were the Gates-MacGinitie Reading Tests-4 (Gates-MacGinitie; MacGinitie, MacGinitie, Maria, & Dreyer, 2000) and the Reading Comprehension subtest of the Wechsler Individual Achievement Tests-III (WIAT-III; Wechsler, 2009). Both tests (similar to the NT tests) were administered by testers "blind" to students' group affiliations. On the Gates-MacGinitie, students read a series of 11 passages, which differed by grade level and represented narrative and informational texts in roughly similar proportions. There were 3 to 6 multiple-choice questions per passage, totaling 48 questions. They required factual recall, inferential reasoning, and an integration of content. Few of them, however, necessitated paragraph- or passage-level summarizing. This FT test was administered pre- and post-treatment in a small-group (6-8 students) or whole-class format in accordance with the test manual. Sample-based Cronbach's alpha was .71 and .75 at pre- and post-treatment for students in grade 3; .77 and .80 for students in grade 5.

The RC subtest of the WIAT-III was administered individually to students at pre- and post-treatment. Students read a set of three mostly narrative passages (per grade level band) and answered open-ended questions about them. They were permitted to view the passages as they answered the questions, most of which required children to recall factual or literal information rather than to summarize text and make inferences about it. Sample-based reliabilities were not obtained because third and fifth graders were assessed on different items. The manual states internal reliabilities of .82 and .91 for students in grades 3 and 5, respectively.

Procedures

Test training and testing fidelity. The two project coordinators explained the purpose and nature of the tests to the RAs and then carefully instructed them on how to administer each test. As part of their training, the RAs were assigned a partner with whom they were expected to practice test administration for 6 hours. Training for post-treatment testing was an abbreviated version of pretreatment training because the RAs by then had become experienced in administering all but three measures, which were given only following study completion.

One week after test training, the RAs were required to reach a 90% criterion for the administration and scoring of each test. Their performance was evaluated by the project coordinators who used a checklist tailored to each test. The 90% criterion had to be met before the RAs were permitted to test participants. If an RA failed to meet criterion, she or he was required to complete further practice and repeat the administration or scoring of the test in question. The RAs had to meet the same 90% criterion for test administration and scoring before post-treatment data collection.

Tutor training and tutoring fidelity. Tutor training was conducted in three half-day sessions. Project coordinators described and provided a rationale for each intervention component, demonstrated the instructional activities, and engaged the RAs in limited role-playing with feedback. Following the three half-day sessions, the RAs practiced using both COMP and [WM]COMP tutoring protocols with another RA for 6 hours. Then they “tutored” a project coordinator who pretended to be a student. They had to display at least 90% adherence to the protocols for both versions of tutoring. Failing this, they engaged in more practice, and their tutoring fidelity was evaluated again. Three of 13 RAs required a second evaluation. Each RA was assigned one or more student pairs from both treatment conditions. Once tutoring commenced, the RAs and project staff met weekly for 1 hour to discuss upcoming lessons and whatever tutoring difficulties they were experiencing.

Project coordinators formally observed the tutors once at the start, middle, and end of the 14-week intervention. On each of these three occasions, the RAs were rated in accordance with a 33-item checklist. In aggregate, the checklist items described expected tutoring behavior: adhering to the instructional script, using standard correction procedures, making appropriate use of tutoring materials, and completing lessons. For each item, the RAs earned 1 point by demonstrating the specified tutoring behavior at least 75% of the time; 0 points if the behavior was demonstrated less frequently. An RA's total score (percentage of adherence) was calculated by dividing the number of items assigned one point by the total number of items. Tutors were also assigned a qualitative rating on three aspects of their performance: pacing of instruction, using correction procedures, and encouraging positive academic and social behavior. Each of these dimensions was rated 1 (highest) to 3 (lowest). Hence, across the three dimensions, "3" was the most positive score, "9" was the least positive score.

The more experienced of the two project coordinators conducted 28 of the 36 observations of the RAs' tutoring sessions—11 of 12 at Time 1, 8 of 12 at Time 2, and 9 of 12 at Time 3. The second project coordinator completed the remaining observations. Because the more experienced coordinator conducted nearly 80% of the observations, and because of logistical concerns, no inter-rater agreement between the two was obtained.

Data Entry and Data Analysis

Data were double-entered by staff members and a small group of RAs. Discrepancies were resolved by staff who returned to the completed test protocols as necessary. Data analysis was conducted to answer our two primary research questions: Do students who receive COMP show superior RC performance after treatment than control students? Do students who receive [WM]COMP show superior RC performance after treatment than controls? We answered these questions by analyzing data (separately for third and fifth graders) from an NT measure of knowledge acquisition, and from NT and FT tests of RC. We were not interested in formally (statistically) comparing effects across grades, but only in estimating them separately to provide a more in-depth understanding of treatment effects in each grade.

Because neither of our FT measures was closely aligned to the skills and strategies taught in the treatment conditions (a point to which we return), pre- and post-treatment factor scores of FT RC measures were calculated for further analysis. (Because these factor scores were weighted equally, they were technically composite scores.) Next, measures of NT knowledge, NT RC, and FT RC were examined by grade to detect distributional irregularities or extreme values. Several measures appeared to depart from normality and were flagged for transformation should model residuals suggest the necessity to do so. In addition, one third-grade student had a pretreatment FT RC score of -3.34 *SD* below the mean. This score was winsorized to the next closest value, -2, to reduce its influence on model estimates.

The structure of the data was different in the treatment arm than the control arm. In the treatment arm, students were cross-classified in classrooms and tutoring pairs that were nested in schools. In the control arm, students were nested in classrooms that were nested in schools. Dependency in the data from our partially nested design was accounted for in the multilevel model according to Sterba (2015). Specifically, we ran unconditional models for each outcome that included a school random effect, classroom random effect, and pair random effect (for students in COMP and [WM]COMP). In addition, level-1 residuals were allowed to vary by condition to account for possible heterogeneity across conditions. Standard errors were corrected for small numbers of clusters by applying the Kenward-Roger adjustment as recommended by Baldwin, Bauer, Stice, and Rohde (2011). Random effects with ICCs of 0 were removed prior to estimating the final model to avoid unnecessarily reducing power (Baldwin et al., 2011).

The final model for each outcome contained the following covariates: pretreatment score on corresponding outcome measure to reduce error variance; ITBS RC raw score to account for the stratified randomization; D_COMP, a dummy variable comparing COMP to controls; D-[WM]COMP, a dummy variable comparing [WM]COMP to controls; and the necessary random effects as determined by non-zero ICCs. ITBS RC scores were correlated with other pretreatment measures at <.60, meaning that extreme collinearity was not present. All models were run in Stata/SE 14.1 with the *mixed* command.

Results

Table 3 shows the fidelity with which the RAs conducted tutoring. "Percentage of adherence" refers to the average proportion of required tutoring behaviors displayed by the tutors during the three times they were observed. The proportion of adherence across the three time points and two treatment conditions ranged from 91% to 99%. Quality ratings of RAs' tutoring behavior were also positive (see table).

To better understand treatment and control groups' reading instruction, we determined the number of students who missed some portion of classroom instruction due to their participation in our tutoring program, and the number of students who got additional reading instruction(that is, instruction beyond what they got in their classrooms and from our tutoring). Data in Table 4 reveal that a large percentage of COMP and [WM]COMP students missed general reading instruction. The most frequently missed instruction was described by teachers as "RTI" or "intervention." Table 4 also shows that a sizable minority of children received instruction in addition to classroom and project-related instruction. It is difficult to know how these circumstances may have affected our results. On the one hand, treatment effects may have been attenuated because some treatment students did not receive a full dose of classroom instruction. On the other hand, treatment effects may have been inflated to the extent that students in one or both tutored groups participated disproportionately in additional school-based (and presumably efficacious) instruction when compared to controls.

The three study groups' pre- and post-treatment means and *SDs* are displayed by grade in Table 5 to help make model results more understandable. Level 1 residuals from final models were checked for normality, extreme values, and homoscedasticity. In no case did residual checks suggest that model modifications were necessary. Results of the fixed effects from the final multilevel models are in Table 6 and are discussed first for NT knowledge acquisition, then NT RC, and finally FT RC. A random effect for school, classroom, and student pair was included as long as the ICC was > 0. Otherwise, it was omitted. Random effects from the final models are not reported due to space constraints.

Results for NT knowledge acquisition are displayed in the top section of Table 6. Controlling for pretreatment vocabulary knowledge and reading comprehension skill, students in grade 3 in the COMP and [WM]COMP conditions significantly outperformed controls, B = 3.26, *SE* = 0.64, *p* < .001, and B = 4.91, *SE* = 0.68, *p* < .001, respectively. Similarly, in grade 5, students in the COMP and [WM]COMP conditions significantly outperformed controls, B = 3.89, *SE* = 0.80, *p* < .001, and B = 4.30, *SE* = 0.75, *p* < .001, respectively. Effect sizes (Hedges *g* with small-sample correction; U.S. Department of Education, 2013) calculated from these model coefficients were large: *ES* = 1.36 and 1.90 in grade 3 for the COMP vs. control and

[WM]COMP vs. control contrasts, respectively; and *ES* = 1.46 and 1.60 in grade 5 for the same contrasts (see Figure 1).

Results of the final multilevel models for NT RC are in the middle section of Table 6. Controlling for pretreatment NT RC and the stratification variable, performance on the NT RC measure was not significantly better for grade 3 students in COMP vs. controls, B = 0.15, *SE* = 0.65, *p* = .817, but was significantly better for grade 3 students in [WM]COMP vs. controls, B = 1.91, *SE* = 0.69, *p* = .009. Hedges *g* for COMP vs. controls and [WM]COMP vs. controls were 0.05 and 0.79, respectively (see Figure 1). In grade 5, students in neither treatment significantly outperformed control students on NT RC, B = 1.60, *SE* = 0.83, *p* = .068 for COMP vs. controls and B = 1.06, *SE* = 1.02, *p* = .312 for [WM]COMP vs. controls. However, Hedges *g* suggests that the treatments produced educationally meaningful effects (*ES* cutoff = 0.25; U.S. Department of Education, 2013), *ES* = 0.52 for COMP vs. controls and *ES* = 0.31 for [WM]COMP vs. controls (see Figure 1).

Results of the final multilevel models for FT RC are in the bottom section of Table 6. Model coefficients and Hedges *g* provide evidence that, controlling for beginning RC skill, students in the grade 3 treatment conditions did not outperform controls on measures of FT RC, B = -0.17, *SE* = 0.26, *p* = .509, *ES* = -0.16 for COMP vs. controls, and B = 0.08, *SE* = 0.27, *p* = .762, *ES* = 0.08 for [WM]COMP vs. controls. Similarly, students in the grade 5 treatment conditions did not demonstrate significantly stronger performance than control students, B = 0.22, *SE* = 0.23, *p* = .361, *ES* = 0.21 for COMP vs. controls, and B = 0.36, *SE* = 0.27, *p* = .188 for [WM]COMP vs. controls, although an *ES* of 0.33 for [WM]COMP vs. controls suggests a small but educationally meaningful treatment effect (see Figure 1).

Figure 2 shows descriptive information from the post-treatment NT main idea and inferences test by item type (factual vs. inferential vs. main idea), treatment condition, and grade. Each symbol represents the proportion of students in a condition correctly answering a given item type. The item types are ordered by their presumed lesser-to-greater difficulty: factual, text-based inference, elaborative inference, and main idea. The texts are presented in the sequence in which they were used: Dean Kamen (text 1), Paper Clip (text 2), Coretta Scott King (text 3), and Civil Rights Act (text 4). We examined these graphs to gain insight into the functioning of item types by treatment condition.

In grade 3 (bottom graph in Figure 2), the two tutored groups consistently outperformed controls on the main idea items (see Main, 1-4), and a bit less consistently on elaborate inference items (see Elab, 1-4). They and control students performed similarly on text-based inference items (Text, 1-4) and factual items (Fact 1-4). The [WM]COMP group tended to do better than the COMP group across item types. Surprisingly, factual items seemed more difficult than text-based inference items, judging from the averaged proportion of items answered correctly by the study groups in grade 3.

Students' performance in grade 5 (top graph in Figure 2) was different from those in grade 3 in several respects. The tutored groups outperformed controls on every factual item and text-based inference item, and did better than controls on a majority of elaborate inference items and main idea items. In contrast to grade 3, the COMP group appeared consistently stronger than [WM]COMP. The figure also suggests that there was a wider range of performance in grade 3 than in grade 5.

Discussion

The purpose of this experimental study was to develop a relatively comprehensive tutoring program that would improve at-risk students' RC in informational texts. Toward this end, we conducted a 14-week field trial in 2015-2016 that involved 50 classroom teachers, 15 tutors, and 120 students drawn in equal proportions from grade 3 and grade 5 in 13 schools in a large urban school district. At the start of the study, these students' word reading was in the average range, whereas their RC was considerably poorer. They were randomly assigned in equal numbers to two tutoring conditions and a control group. Both tutored groups engaged in a base RC program that addressed before-reading activities (e.g., discussing title, text features, and vocabulary words and developing guiding questions) and after-reading activities (e.g., retelling important facts, developing main idea statements, and answering factual and inferential questions). One of the tutored groups also participated in explicit WM training, which was embedded in the RC instruction.

Before discussing our findings, several study limitations deserve mentioning because they represent potential threats to the validity of our effort. First, we had only 18 to 20 students in each of the three study groups at grade 3 and grade 5—a number that may have adversely affected our capacity to find a greater number of statistically significant effects. More importantly, a larger number of students would have permitted a comparison between the two treatments, in addition to the comparisons of each to controls, because of the greater power produced to correct alpha for multiple comparisons.

Second, Table 3 shows that a sizable minority of students (a) missed reading instruction in their general program because of their participation in our tutoring and/or (b) obtained reading instruction from others who were not our tutors or their classroom teachers. These facts potentially complicate interpretation of our findings. For example, the reading performances of children receiving additional small-group instruction may reflect multiple treatment effects. A third limitation, mentioned earlier, was that we were unable to estimate inter-rater agreement for the two project coordinators who observed the fidelity with which the RAs conducted tutoring. Fourth, our tutoring program was 2:1. Arguably, this is a ratio that doesn't generalize to many districts in which tutored groups tend to be considerably larger.

There are at least two more limitations that are more limitations of study presentation than of study execution. Constraints on the length of this manuscript precluded a description of WM training effects on WM performance. A description of our WM measures and outcomes will appear elsewhere. Similarly, we did not mention, let alone explain, how and why we developed our own informational texts. We will provide such a description in the future.

These considerations notwithstanding, study results indicated that students in the two tutoring conditions tended to outperform controls (more so in grade 5 than in grade 3) on NT measures but not on FT measures (see Figure 1). On the NT test of knowledge acquisition, both COMP and [WM]COMP groups statistically significantly outperformed controls in grades 3 and 5 (*ES* ranging from 1.36 to 1.90). On the NT test of RC, results were mixed. At grade 3, the [WM]COMP group reliably outperformed controls (*ES* = 0.79), whereas COMP and control students performed equally, At grade 5, neither treatment group performed differently from controls, although the COMP versus control contrast approached significance (*p* = .067; *ES* = 0.52). There were no statistically significant between-group differences in either grade on our FT composite index of RC.

We suspect that many in the educational research community would regard these results as mostly unimportant, and would characterize our effort as a study of null effects. We base this suspicion on what we believe is a consensual view that devalues NT measures and values FT ones. Consider, for example, the following from the What Works Clearinghouse (WWC), which describes its purpose as conducting reviews of research "to provide educators with the information they need to make evidence-based decisions" by focusing on results from high-quality studies (https://ies.ed.gov/ncee/wwc/). The WWC wrote, "To be eligible for review, [a study] outcome must (a) [have] face validity and reliability, (b) not be over-aligned with the intervention, and (c) be collected in the same manner for both intervention and comparison groups" (WWC, 2014, p. 16). On page 17 in the same document, the WWC clarified what is meant by the term, *over-aligned*:

> When outcome measures are closely aligned with or tailored to the intervention, the study findings may not be an accurate indication of the effect of the intervention. For example, an outcome measure based on an assessment that relied on materials used in the intervention condition but not in the comparison condition likely would be judged to be over-aligned.

Although the WWC did not discuss NT measures by name, we believe (a) it is describing them as over-aligned tests of learning acquisition that unfairly advantages children in the treatment group and (b) it is discouraging their use in favor of more distal measures. Again, this preference for FT tests is a bias that we believe is shared by many. Indeed, we're certain that many scholarly journals would have rejected this paper for publication because of our failure to demonstrate reliable between-group differences on our FT composite index.

It is time to reconsider the lowly reputation of tests of NT. Development of instructional programs that accelerate student learning is an iterative undertaking. It requires conducting an initial version, evaluating its effects, modifying it in light of what is learned, implementing the revised program, and so forth. Student performance on our NT measures greatly informs this process of development precisely because the NT tests are aligned to our program. (Because these NT tests require children to apply taught skills and strategies to unfamiliar content, they assess *transfer*, not merely *learning acquisition*.) Data in Figures 1 and 2 help us know whether and to what extent children learned what we intended them to learn, and which skills and content were taught more successfully than others. (We believe readers would be similarly interested in these results.)

By contrast, test items on the FT measures—the Gates-MacGinitie (MacGinitie et al., 2000) and the WIAT III (Wechsler, 2009)—correspond much less well to our program content. We know this because for each item on the two FT tests, we informally examined whether one or more of our eight skills and strategies (i.e., the before-reading, after-reading, and supplementary activities) would have helped study participants respond correctly. Across the two FT tests and two grade levels, only 1 of 8 skills/strategies would have been helpful on more than 50% of test items; 5 of 8 would have been helpful when answering 10% to 50% of test items; and 2 of 8 would have been helpful on less than 10% of the test items. One might argue that one or more of our 8 skills and strategies are of dubious worth and don't deserve to be part of the FT measures, but our skills/strategies are all evidence-based.

We are not saying program developers should rid FT measures from their test batteries. Such tests provide unique and important information; they have a role to play in program evaluation. But the same must be said for NT measures. They may be especially useful in regard to RC. Because of its multi-dimensionality, program developers and test developers have often operationalized the RC construct in numerous and very different ways. For example, RC programs may emphasize inference-making, whereas RC tests of FT may focus on factual knowledge. Given the frequent disjunction between RC programs and tests, one may wonder if program developers should design instructional routines addressing the content of the Gates-MacGinitie, WIAT II, or other well-known FT measures. We believe this would be a mistake.

RC programs should be based on science and on the developers' knowledge of children, teachers, schools, and communities. If program development was driven by FT test content, RC programs might enhance performance on such measures at the expense of failing to strengthen children's RC more broadly. In short, NT and FT tests should be seen in principle as having equally important roles and a complementary relationship to each other.

However reasonable this last sentence may sound, it begs an important question: If NT measures are to be taken more seriously, then how should we define the "N" and the "T"? Consider that RC may be defined as recalling factual information or making main idea statements or constructing inferences; that test stimuli may be presented visually or orally; that children may be required to respond orally or in writing to multiple-choice or open-ended questions; and that tests may be timed or untimed, administered individually or in small or large groups. On which of these dimensions should program developers attempt to match their instruction and outcome measures? On all? On a few? And if on a subset, on what basis would it be chosen? More generally, there is an infrequently recognized tension when attempting to create or select a test that is sensitive to the objectives of instruction and that explores a student's capacity to generalize what is learned from the instruction to new contexts. Currently, there is little guidance available on how to proceed.

Whereas much of this discussion has focused on how and why our treatment groups performed differently on NT and FT measures, there was also performance variability across grades and treatment conditions. Rather than attempt to explain all of this inconsistency (of which we are incapable), we wish to draw attention to the dissimilarities in performance between just the COMP groups at grades 3 and 5 on our NT and FT measures of RC (see Figure 1). Results indicated that the COMP treatment failed to improve most third-grade participants' RC in an absolute sense and relative to the COMP group's performance at grade 5 (see Figure 1). We believe there were at least three reasons for this. First, the informational texts we wrote were too difficult for weak readers in third grade. Second, we attempted to teach too many strategies and, as a result, taught few of them well. Third, we did not provide sufficient opportunity or encouragement for the two children in each tutored group to interact meaningfully with each other. We have reason to believe that peer-mediated academic work, when appropriately designed, is a unique and empowering dimension to instructional programs. These last two points, we believe, also speak to our instructional approach in grade 5.

It is likely that these reasons in combination contributed to the COMP treatment's less-than-satisfactory influence on children's RC, especially in third grade. Further, it weakened our examination of the importance of our second treatment condition, [WM]COMP. As we write, we have preliminary results from a follow-up study conducted in grades 3, 4, and 5 in 2016-2017. Whereas the follow-up study had the same three groups as the 2015-2016 study, we developed more appropriate informational texts for third graders in the follow-up study. We also produced a more focused instructional program for COMP and [WM]COMP, emphasizing main idea identification and in-text and out-of-text inference-making; a greater number of opportunities for peers to interact meaningfully with each other; and a stronger set of NT measures of RC. Findings from this follow-up effort are more encouraging, which permits us to make a final point. Serious program development is an iterative process that requires resources and a willingness to learn from mistakes. We are hopeful we will eventually produce an efficacious RC program for children and teachers before our resources and patience run out.

References

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16,* 149-165. doi: 10.1037/a0023464

Becker, W. C., Englemann, S., Carnine, D., & Rhine, R. (1981). Direct instruction models. In R. Rhine (Ed.), *Encouraging change in American schools: A decade of experimentation* (pp. 45-83). New York: Academic Press.

Berninger, V., & Richards, T. (2010). Inter-relationships among behavioral markers, genes, brain, and treatment in dyslexia and dysgraphia. *Future Neurology, 5,* 597-617. doi: https://doi.org/10.2217/fnl.10.22

Carretti, B., Caldarola, N., Tencati, C., & Cornoldi, C. (2014). Improving reading comprehension in reading and listening settings: The effect of two training programmes focusing on metacognition and working memory. *British Journal of Educational Psychology, 84*, 194-210. doi: 10.1111/bjep.12022

Compton, D. L., Miller, A. C., Gilbert, J. K., & Steacy, L. M. (2013). What can be learned about the reading comprehension of poor readers through the use of advanced statistical modeling techniques? In L. E. Cutting, B. Miller, & P. McCardle (Eds.), *Unraveling the behavioral, neurobiological, & genetic components of reading comprehension* (pp. 135-147). Baltimore: Brookes.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review, 3,* 422-433. doi: https://doi.org/10.3758/BF03214546

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211-245. doi: http://dx.doi.org/10.1037/0033-295X.102.2.211

Fuchs, D., Elleman, A. M., Fuchs, L. S., Peng, P., Kearns, D. M. Compton, D. L., …, Miller, A. C. (2016). *A randomized control trial of explicit instruction with and without cognitive training to strengthen the reading comprehension of poor readers in first grade.* Manuscript submitted for publication.

Garcia-Madruga, J. A., Elosua, M. R., Gil, L., Gomez-Veiga, I., Vita, J. O., Orjales, I., Contreras, A., Rodriguez, R., Melero, M. A., & Duque, G. (2013). Reading comprehension and working memory's executive processes: An intervention study in primary school students. *Reading Research Quarterly*, *48*, 155-174. doi: 10.1002/rrq.44

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research*, *71*, 279-320. doi: 10.3102/00346543071002279

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *Iowa Tests of Basic Skills.* Itasca, IL: Riverside.

Kintsch W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Current issues in reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Lawrence Erlbaum Associates.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163-182.

MacGinitie, W., MacGinitie, R., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests* (4th ed.). Itasca, IL: Riverside Publishing Company.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*, 270-291. doi: http://dx.doi.org/10.1037/a0028228

Meneghetti, C., Carretti, B., & De Beni, R. (2006). Components of reading comprehension and scholastic achievement. *Learning and Individual Differences, 16*, 291-301. doi: https://doi.org/10.1016/j.lindif.2006.11.001

Nation, K. (2005). Children's reading comprehension difficulties. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 248-265). Oxford, England: Blackwell.

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

O'Connor, R., & Fuchs, L. S. (2013). Responsiveness to intervention in the elementary grades: Implications for early childhood education. *Handbook of response to intervention (RTI) in early childhood education* (pp. 41-56). Baltimore: Brookes.

Palladino, P., Cornoldi, C., De Beni, R., & Pazzaglia, F. (2001). Working memory and updating processes in reading comprehension. *Memory & Cognition, 29*, 344-354. doi: https://doi.org/10.3758/BF03194929

Schwaighofer, M., Fischer, F., & Buhner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist, 50*, 138-166. doi: http://dx.doi.org/10.1080/00461520.2015.1036274

Sterba, S. K. (in press). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research.*

Swanson, H. L., & Howell, M. (2001). Working memory, short-term memory, and speech rate as predictors of children's reading performance at different ages. *Journal of Educational Psychology*, *93*, 720-734. doi: http://dx.doi.org/10.1037/0022-0663.93.4.720

Swanson, H. L., & O'Connor, R. (2009). The role of working memory and fluency practice on reading comprehension of students who are dysfluent readers. *Journal of Learning Disabilities, 42*, 548-575. doi: 10.1177/0022219409338742

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of Word Reading Efficiency* (2nd ed.). Austin: Pro-Ed.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: an examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*, 1038-1066. doi: http://dx.doi.org/10.1037/0033-2909.133.6.1038

Wechsler, D. (2009). *Wechsler Individual Achievement Test* (3rd ed.). San Antonio: Psychological Corporation.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence* (2nd ed.). San Antonio: NCS Pearson.

What Works Clearinghouse. (2014). Procedures and standards handbook, version 3.0. Washington, D.C.: Institute for Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Protocols#procedures

Table 1

Descriptive Statistics for Teacher Demographics by Grade

	Grade 3 (*n* = 28)		Grade 5 (*n* = 22)	
Variable	*f*	%	*f*	%
Female	28	100.00	18	81.82
African American	2	7.14	3	13.64
Caucasian	25	89.29	18	81.82
Hispanic	1	3.57	1	4.55
Highest Educational Degree				
B.S./B.A.	5	17.86	6	27.27
B.S./B.A. +	1	3.57	1	4.55
M.Ed./M.S.	16	57.14	9	40.91
M.Ed./M.S. +	5	17.86	4	18.19
Ed.S.	1	3.57	1	4.55
Ed.D/Ph.D.	0	0.00	1	4.55
ELL Certification	14	50.00	4	18.18
Reading Certification	6	21.43	11	50.00
Special Education Certification	2	7.14	0	0.00
	M	*SD*	*M*	*SD*
Years in current position	10.29	6.58	10.27	9.27
Years in teaching profession	2.79	2.34	3.25	2.59

Table 2

Student Demographics and Performance on Screening Measures by Grade

	Grade 3 (*n* = 59)		Grade 5 (*n* = 57)	
Variable	*f*	%	*f*	%
Female	27	45.76	30	52.63
African American	11	18.64	34	59.65
Caucasian	15	25.42	19	33.33
Hispanic	27	45.76	1	1.75
Other	6	10.16	3	5.26
Free/Reduced Price Lunch	56[a]	96.55	51[b]	91.07
Individualized Education Plan	1	1.69	6	10.53
	M	*SD*	*M*	*SD*
ITBS Reading Comprehension DSS	166.19	10.42	184.88	14.73
TOWRE-2 Sight Word Efficiency SS	101.10	6.23	94.00	8.53
WASI-2 Matrix Reasoning T-score	44.56	7.72	42.30	7.87
WASI-2 Vocabulary T-score	46.42	6.43	47.11	6.69

Note. ITBS is Iowa Test of Basic Skills (Hoover, Dunbar, & Frisbie, 2003). DSS is developmental standard score; the following DSSs represent typical performance at the end of grade 2, 3, 4 and 5, respectively: 168, 185, 200, and 214. TOWRE-2 is Test of Word Reading Efficiency-2 (Wagner, Torgesen, & Rashotte, 2012). WASI-2 is Wechsler Abbreviated Scale of Intelligence (Wechsler, 2011). T-scores have *M*s of 50 and *SD*s of 10.

[a]*n* = 58 on lunch status variable due to missing data. [b]*n* = 56 on lunch status due to missing data.

Table 3

Quantitative and Qualitative Fidelity of Treatment Data by Condition and Time Point

	COMP			[WM] COMP		
Fidelity Index	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
% Adherence	99.00 (1.64)	93.00 (5.65)	98.10 (3.14)	91.00 (9.00)	97.00 (3.00)	97.50 (3.54)
Quality Rating[a]	3.00 (0.45)	4.00 (0.82)	3.10 (0.32)	4.00 (1.00)	4.00 (1.00)	3.00 (0.00)

[a] Sum of three ratings on a 1 (highest) to 3 (lowest) scale. 3 is best possible score; 9 is worst possible score.

Table 4

Numbers of Students Who Missed Reading Instruction or Received Additional Reading Instruction by Condition across Grades

	COMP (*n* = 40)	[WM]COMP (*n* = 38)	Control (*n* = 38)
Missed Reading Instruction because of Treatment	12 (30.00%)	18 (47.37%)	n/a
Type of Missed Instruction			
Balanced Literacy	2	1	
Centers	0	4	
Independent Work	1	6	
PALS/Guided/Shared Reading	2	3	
RTI/Intervention	8	7	
Skills Practice	1	5	
Story Time/Whole Group Reading	0	2	
Writing	0	3	
Other	0	4	
Reading Instruction in Addition to Classroom and Project-Related Instruction	11 (27.50%)	6 (15.79%)	5 (13.16%)
Type of Additional Instruction[a]			
After School Tutoring	3	2	0
ELL Tutoring	2	1	0
Individual/Small Group Tutoring	3	2	3
Occupational Therapy	0	0	1
Other	3	2	2

Note. COMP is reading comprehension training. [WM]COMP is reading comprehension training with embedded working memory training. Control is business-as-usual school-based reading instruction.

[a]These categories are not mutually exclusive. Students may be counted in more than one category.

Table 5

Means (and SDs) for Pre- and Post-treatment Measures by Grade and Condition

	Grade 3			Grade 5		
Variable	COMP (*n* = 20)	[WM]COMP (*n* = 19)	Control (*n* = 20)	COMP (*n* = 20)	[WM]COMP (*n* = 19)	Control (*n* = 18)
Pre WASI-2 Vocabulary (proxy for general knowledge)	20.50 (3.27)	18.53 (2.89)	21.10 (4.29)	26.50 (4.02)	26.53 (4.50)	24.44 (3.58)
Post NT Knowledge Acquisition	13.75 (2.29)	14.95 (2.68)	10.20 (2.40)	16.15 (2.28)	16.63 (2.31)	11.89 (2.93)
Pre NT Reading Comprehension	8.30 (3.06)	7.79 (2.78)	7.65 (3.22)	10.75 (2.84)	10.58 (3.20)	10.19 (3.29)[a]
Post NT Reading Comprehension	9.85 (2.92)	11.26 (2.13)	9.20 (2.57)	13.05 (2.14)	12.47 (3.01)	11.11 (3.71)
Pre FT Reading Comprehension	0.06 (1.03)	0.03 (0.96)	-0.02 (0.86)	0.14 (0.90)	0.14 (1.09)	-0.33 (0.99)[b]
Post FT Reading Comprehension	-0.09 (1.02)	0.12 (0.96)	-0.02 (1.05)	0.14 (0.76)	0.27 (0.88)	-0.44 (1.23)

Note. COMP is reading comprehension training. [WM]COMP is reading comprehension training with embedded working memory training. Control is business-as-usual school-based reading instruction. Pre is pretreatment. WASI-2 is Wechsler Abbreviated Scale of Intelligence (Wechsler, 2011). Post is post-treatment. NT is near-transfer. FT is far-transfer.

[a] *n* = 16 due to staff error. [b] *n* = 17 due to administration error.

Table 6

Multilevel Model Fixed Effects Results across Post-treatment Outcomes by Grade

	Grade 3 (n = 59)				Grade 5 (n = 57)			
Outcome/Fixed Effect	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*
Near Transfer Knowledge								
Intercept, γ_{00}	2.81	1.95	1.45	.156	3.51	2.16	1.63	.103
Pretreatment WASI-2 Vocabulary, γ_{01}	0.12	0.08	1.58	.121	0.21	0.08	2.81	.005
Pretreatment ITBS Reading Comprehension, γ_{02}	0.36	0.07	4.87	<.001	0.17	0.07	2.57	.010
D_COMP, γ_{03}	3.26	0.64	5.08	<.001	3.89	0.80	4.84	<.001
D_[WM]COMP, γ_{04}	4.91	0.68	7.24	<.001	4.30	0.75	5.72	<.001
Near Transfer Comprehension								
Intercept, γ_{00}	2.96	1.04	2.85	.006	4.66	1.76	2.65	.011
Pretreatment Near-Transfer Reading Comprehension, γ_{01}	0.38	0.09	4.09	<.001	0.41	0.13	3.27	.002
Pretreatment ITBS Reading Comprehension, γ_{02}	0.25	0.08	3.14	.003	0.13	0.09	1.47	.149
D_COMP, γ_{03}	0.15	0.65	0.23	.817	1.60	0.83	1.92	.068
D_[WM]COMP, γ_{04}	1.91	0.69	2.78	.009	1.06	1.02	1.03	.312
Far Transfer Comprehension								
Intercept, γ_{00}	-0.57	0.46	-1.25	.214	-0.41	0.48	-0.86	.396
Pretreatment Far-Transfer Reading Comprehension, γ_{01}	0.67	0.12	5.45	<.001	0.64	0.10	6.24	<.001

Pretreatment ITBS Reading Comprehension, γ_{02}	0.04	0.03	1.42	.160	0.01	0.02	0.57	.568
D_COMP, γ_{03}	-0.17	0.26	-0.66	.509	0.22	0.23	0.93	.361
D_[WM]COMP, γ_{04}	0.08	0.27	0.30	.762	0.36	0.27	1.36	.188

Note. WASI-2 is Wechsler Abbreviated Scale of Intelligence (Wechsler, 2011). ITBS is Iowa Test of Basic Skills (Hoover, Dunbar, & Frisbie, 2003). D_COMP is dummy variable comparing reading comprehension training to control. D_[WM]COMP is dummy variable comparing reading comprehension training with embedded working memory training to control. Necessary random effects were included in each model but not presented here for the sake of space. Grade 5 had only 55 and 56 students, respectively, in the Near Transfer Comprehension and Far Transfer Comprehension models due to missing data.