Developing an Informational Text Comprehension Intervention for Struggling Readers in Third Grade

By

Meagan E. Walsh

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Special Education

June 30, 2020

Nashville, Tennessee

Approved:

Douglas Fuchs, Ph.D.

Lynn S. Fuchs, Ph.D.

Marcia A. Barnes, Ph.D.

Amy Elleman, Ph.D.

To my daughters, Winter, Meadow, Rhapsody, Arabesque, and Quintana, my biggest fans and toughest critics

ACKNOWLEDGEMENTS

The research described in this report was conducted at the National Center on Accelerating the Academic Achievement of Students with Severe and Persistent Learning Disabilities, which was funded by Grant No. R324D130003 from the National Center on Special Education Research, Institute for Educational Sciences. The Principal Investigators were Doug Fuchs and Lynn Fuchs. My doctoral program of study was funded by a Leadership Training Grant awarded to the Department of Special Education, Peabody College, Vanderbilt University (Principal Investigator: Joseph Wehby).

I am deeply indebted to Dr. Doug Fuchs for his guidance and support throughout my Ph.D. program. His advice on all matters research and career has been unapparelled. I am especially grateful for the freedom he afforded me to create, explore, and study while developing and iterating the third grade curriculum. Thank you for taking a chance on this special educator, for sharing your research and insight with me, and for always encouraging me to put "the work" first, that is, the rigorous and difficult task of developing intensive interventions for students with and at-risk for disabilities.

For me, this work began in the classroom and continues to be motivated by my students. Dear students, please know that I will never stop fighting for you. Thank you, especially, to Jasmine, Michael, and Triston. This was the intervention you needed. I'm sorry it took me so long.

Throughout the Ph.D. process I was sustained by a cohort of exceptional women. Thank you to Kathleen Zimmerman, Emily Weaver, Sarah Krowka, Emily Quinn, Jodi Heidlage, and Casey Chauvin for your encouragement and support. Your strength, intelligence, wit, and dedication continue to inspire me.

The specific project described in this dissertation would not have been possible without the dedication and effort of several individuals. I am especially thankful to the project coordinator on this study, Pamela Lehman, my doctoral colleague, Samuel Patton, and the incomparable Lynn Davies for their tireless support. I am also thankful to each of the research assistants who worked on this project with us—Sarah Raider, Michelle Nagy, Hope Wiggs, Harrison Baker, and Sydney Padula. All, your feedback and attention to detail made the intervention what it is.

The version of the intervention described in this dissertation was the last in a long line of intervention iterations each supported by an equally dedicated team of project coordinators, doctoral students, and research assistants. I am thankful for each member of the Fuchs research team with whom I had the pleasure of working. However, I hold a special reserve of gratitude in my heart for Loulee Yen Haga who welcomed me to research and the Nashville community at large. Loulee, although you were taken too soon, your energy and light continue to guide my efforts. May you rest in joy.

Finally, this accomplishment is shared, in full, with my husband, Dan Walsh, who sacrificed his own goals to make this dissertation a reality. He stood by me through a myriad of dead ends and dark days. Together we weathered tornados, pandemics, and five daughters. Ours was an unusual path. Thank you for sharing the burden, for giving me space, and believing the best, always. We made it. Onward.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Prior to third grade, reading instruction is primarily focused on helping students to acquire foundational reading skills (i.e. phonology, decoding, sight words, etc.). By the time students encounter the first round of high stakes testing in the spring of the third grade year, they are expected to have transitioned from "learning to read" into "reading to learn." However, proficient reading is not an "inevitable outcome" of proficient word-reading (Snow, 2002). Many students, including those with adequate word-reading abilities, fail to make this transition and subsequently "fail to thrive" in later grades (e.g. Wanzek, Wexler, Vaughn, & Ciullo, 2010). Research suggests that students who fail to meet third-grade proficiency standards in reading are at increased risk for a host of adverse outcomes throughout their academic careers and beyond (e.g. Hernandez, 2011). In principle, an effective third grade reading intervention could help at-risk readers navigate this curricular shift (Gilbert et al., 2013).

Despite the apparent importance of third grade instruction, recent reviews indicate that it has received less attention in reading research compared with other grades (e.g. Wanzek et al, 2016; Scammacca, Roberts, Vaughn, & Stuebing, 2016). Experimental studies of reading interventions that have included third-grade students have reported notably smaller effects for this subgroup of children (e.g. Mastropieri, Bakken, Scruggs, & Whedon, 1996; National Reading Panel [NRP], 2000; Wanzek & Vaughn, 2007). Although research supports the efficacy of word-reading interventions in K-2, these interventions seem to become less effective in third grade (e.g. Lovett, Frijters, Wolf, Steinbach, Sevcik, & Morriss, 2017; Wanzek & Vaughn 2007; Wanzek et al., 2016). Moreover, the robust effects of early word-reading interventions seem to dissipate over time (e.g. O'Connor & Fuchs, 2013). Given the apparent diminishing returns of word-reading interventions, instruction focused on reading comprehension might be more beneficial at third grade.

Reading comprehension interventions, especially those for young and struggling readers, may need to consider the unique demands posed by informational text. Such text has proven especially challenging for many students possibly because it often includes unfamiliar content and text structures (e.g. Cervetti, Janes, & Hiebert, 2009; Gersten, Fuchs, Williams, & Baker, 2001). Also, many young readers and weak readers have knowledge and experiential deficits that make it hard for them to process informational text (Gersten et al., 2001). This difficulty threatens student success in academic content areas (e.g. Gajria, Jitendra, Sood, & Sacks, 2007; Meneghetti, Carretti, & DeBeni, 2006). Thus, it seems that interventions with a focus on comprehending informational texts are necessary to improve outcomes for at-risk third-grade readers.

Reading Comprehension Intervention Research

Reading comprehension instruction has been the subject of intensive focus among researchers since the 1990's (Scammacca et al., 2016). Generally, evidence from extant syntheses of K-12 reading comprehension interventions suggests that these interventions can be effective across the span of reading development (e.g. Mastriopieri et al., 1996; Scammacca et al., 2015). There is general consensus that effective reading comprehension interventions, regardless of grade level, include: explicit strategy instruction, support for the construction of well-integrated situation models, high quality texts, and peer mediation (e.g., Gersten et al., 2001; Shannahan et al., 2010).

Explicit strategy instruction. Since the 1970's, explicit strategy instruction has been the cornerstone of reading comprehension intervention research (e.g. Scammacca et al., 2016). Teaching even a single strategy has been shown to be beneficial (NRP, 2000). Teaching multiple strategies has been shown to be more so (e.g. Gajria et al., 2007; Shannahan et al., 2010). However, "conscious use of strategies comes at a cost" (Elleman & Compton, 2017). Younger readers and struggling readers may have more limited cognitive reserves to devote to strategy use, less sensitivity to detect when strategy use is required, and less developed strategic problem-solving skills (Paris, Lipson, & Wixon, 1983).

Teaching too many strategies, or the wrong strategies, may ultimately detract from true comprehension (e.g. Elleman & Compton, 2017). For this reason, strategy specificity is important to prevent taxing limited cognitive resources unnecessarily (Gersten et al., 2001). Unfortunately, there is no consensus on which strategies in what combinations are most appropriate for a given learner (e.g. NRP, 2000).

Support for construction of well-integrated situation models. Elleman and Compton (2017) argued that researchers and practitioners have neglected the knowledge deficits of young and poor readers in their almost exclusive focus on explicit strategy instruction. They and others have argued that interventions should also focus on knowledge acquisition, creating the context for comprehension. Lack of background knowledge has been shown to contribute to students' inability to generate coherent situation models (Catts & Kamhi, 2017). Failure to construct a situation model is often a symptom of reading disability (Compton et al., 2014). Consequently, there seems to be growing agreement that effective reading comprehension interventions should promote knowledge acquisition and include scaffolded support for student construction of situation models.

High quality texts. Reviews of reading interventions point to the importance of "quality" texts (Shanahan et al., 2010; Gersten et al., 2001). The RAND report (Snow, 2002) called for better texts and better sequences of texts aligned to students' reading levels and interests, the comprehension task at hand, the curriculum, and the wider corpora of texts with which students come in contact. Thus, the evidence indicates that effective interventions should incorporate quality texts that build skill and knowledge simultaneously (Elleman & Compton, 2017).

Peer mediation. Findings from prior syntheses suggest peer learning may also have a facilitative effect on comprehension development by increasing opportunities to respond within the intervention context (e.g. Rosenshine & Meister, 1994). Syntheses indicate that increasing the frequency and meaningfulness of interactions between and among students and teachers is critical when providing intensive intervention and that this can best be accomplished in small collaborative groups (e.g., Gersten et al.., 2001; Webb, 1989).

Measurement issues. Although previous integrative reviews generally conclude that effective comprehension interventions should include some combination of the just-discussed dimensions, such interventions have produced mixed findings. Some of this inconsistency seems attributable to the publication date of reviews. Those published before 2000 tend to report larger effects (cf. Mastropieri et al., 1996; Scammacca et al., 2015). Some of the within-study heterogeneity of intervention effects may be attributed to outcome measures. Specifically, performance on researcher-created measures has generally been stronger than performance on commercially-developed measures (e.g., Clemens & Fuchs, 2020; D. Fuchs et al., 2018; Wanzek et al., 2016). Yet, despite the inconsistency in program effects, views on what makes for an effective reading comprehension intervention have not changed much in the last 25 years (e.g. Rosenshine & Meister, 1994).

The Particular Case of Grade 3

Reading development in third grade. There is reason to believe that third-grade students may be different from older readers in several important ways. First, still-maturing oral language and word-reading abilities may make it difficult for even average third-grade students to derive meaning from connected text (Schwanenglugel et al., 2016). Second, third graders have less background knowledge than older children (Elleman & Compton, 2017). Third, they may have "lower standards of [text] coherence" compared to older peers (McMaster, Espin, & van den Broek, 2014, p.19). This last difference has two implications: With regard to constructing situation models, young and struggling readers tend to make fewer inferences and may even fail to use simple linguistic devices (i.e. anaphora) to build coherent text representations (Elleman, 2017). With regard to strategy use, Cross and Paris

(1988) found that whereas third-grade students might learn and demonstrate a strategy in isolation, they are less aware of the strategy's broader usefulness and are less likely to apply it in context than fifth-grade children.

Perhaps because of these developmental differences, reading comprehension measures have proven less reliable at identifying struggling comprehenders in third grade and in earlier grades. This seems especially true when comprehension difficulty cannot be attributed to word-reading deficits (e.g. Keenan & Meenan, 2014). Without reliable methods of identification, it is difficult to accurately identify the students in need of comprehension intervention in third grade (Fuchs & Fuchs, 2015). Comprehension interventions that are poorly aligned to student needs are likely to be less effective.

Review of third grade comprehension interventions. To better understand the state of comprehension intervention in third grade, Walsh (2020) conducted a meta-analysis of third-grade reading comprehension interventions. The reviewer identified 20 experimental ($m = 10$) and quasi-experimental ($m = 10$) group design investigations that included disaggregated results for struggling third-grade readers. Seven reports were peer reviewed, eleven were dissertations, one was a government report, and one was an unpublished replication of a report published in a book chapter. Studies were conducted between 1983 and 2018. Most ($m = 15$) included some form of explicit strategy instruction, and several ($m = 9$) included peer mediation. Eight investigations were multi-component, combining word-reading and reading comprehension instruction. These 20 studies yielded 54 correlated effect estimates on a variety of researcher-created and commercially-developed measures.

The aggregate effect from these studies was small but significant ($g = 0.16$, robust $SE = 0.08$, $p < 0.05$, 95% CI [0.02, 0.30]). This effect was smaller than the effects reported in other reviews (e.g. Berkeley, Scruggs, & Mastropieri, 2010; Edmonds, et al., 2009) and suggested that reading comprehension interventions may have been somewhat less effective at third grade. Similar to other reviews, however, a smaller effect was reported for commercial measures ($g = 0.08$, robust $SE = 0.05$, 95% *CI* [-0.02, 0.18]); than for researcher-created measures (*g* = 0.51, robust *SE* = 0.23, 95% *CI* [0.06, 0.96]). Also, studies published before 2000 reported larger effects (*g* = 0.39, robust *SE*= 0.13, 95% *CI* [0.14, 0.64]); compared with *g* = 0.06 (robust *SE* = 0.04; 95% *CI* [-0.02, 0.13]) after 2000. Given the large performance gap reported in the literature for students with and at-risk for learning disabilities (Gilmour, Fuchs, & Wehby, 2019), the small and often negligible effect sizes reported on commercial measures and in later studies suggest that third-grade comprehension interventions, as previously studied, have not been effective enough to yield substantive changes in student performance.

Moreover, these studies do not provide evidence that the traditional pillars of comprehension intervention have been consistently effective at third grade. Interventions *without* explicit strategy instruction had a larger, albeit more variable, aggregate effect (*g* = 0.33, robust *SE* = 0.18, 95% *CI* [-0.02, 0.68]); compared to interventions with strategy instruction, *g*=0.13, (robust *SE*=0.05, 95% *CI* [0.03, 0.28]). Similarly, the aggregate effect for interventions incorporating peer mediation was negligible and highly variable (*g* = -0.07, robust *SE* = 0.13, 95% *CI* [-0.32, 0.18]); whereas interventions *without* peer mediation yielded a significant positive effect (*g* = 0.25, robust *SE* = 0.09, 95% *CI* [0.07, 0.43]). Few third-grade studies reported information on the nature, level, or amount of text to which students were exposed. Nor did studies describe the nature of instructional activities meant to support situation model development, with the notable exception of Connor and colleagues (2018), which reported only non-significant, negligible effects for all outcome measures.

There was some evidence from the review, that individual interventions can be differentially effective depending on grade level. Four included studies reported effect sizes separately for more than one grade level (Torgesen et al., 2006; Connor et al., 2018; Fuchs et al, 2018; Walsh, Fuchs, Patton, & Gilbert, 2017). Table 1 displays effect sizes by grade for the various treatment conditions and outcome measures used in each of these studies. Generally, interventions which were effective at one grade level, were not effective at another. Two of the studies compared versions of an explicit strategy intervention (Fuchs et al., 2018; Walsh, et al., 2017). In both, statistically significant effects were reported for fourth-and fifth-grade students, while effects at third grade were generally negligible. Torgesen and colleagues (2006) compared the impact of a multi-component intervention, Failure Free Reading, at third and fifth grade. Although not significant, effects were larger for third grade.

Table 1

Comparison of Reported Comprehension Intervention Effects by Intermediate Grade Level from Multi-Grade Studies.

			Effect Size by Grade		
Study	Treatment	Measure	3rd	4th	5th
Torgesen et al., 2006	Failure Free Reading	Woodcock Reading Mastery Test, Passage Comprehension	0.18		0.02
		Group Reading Assessment and Diagnostic Evaluation	0.35		-0.11
Connor et al., 2018	Enacted Reading Comprehension	Test of Narrative Language Skills	0.10	-0.09	
		Comprehension Monitoring Inconsistency Detection	0.07	-0.09	
		Gates-MacGinite Reading Test	-0.09	-0.08	
		Test of Silent Reading Efficiency and Comprehension	0.04	0.04	
Fuchs et al., 2018	Paired Nonfiction Comprehension	Far Transfer Composite (Gates-MacGinite, Weschler Indivdiual Achievement Test, Woodcock-Johnson, Passage Comprehension)	-0.16		0.21
		Near Transfer Reading Comprehension	0.05		0.52✢
	Paired Nonfiction Comprehension + Embedded Working Memory Training	Far Transfer Composite (Gates-MacGinite, Weschler Indivdiual Achievement Test, Woodcock-Johnson, Passage Comprehension)	0.08		0.31
		Near Transfer Reading Comprehension	0.79**		0.33
Walsh, Fuchs, Patton, & Gilbert, 2017	Individual Nonfiction Comprehension	Far Transfer Composite (Gates-MacGinite, Weschler Indivdiual Achievement Test, Woodcock-Johnson, Passage Comprehension)	0.15	0.53*	
	Individual Nonfiction Comprehension + Embedded Working Memory Training	Far Transfer Composite (Gates-MacGinite, Weschler Indivdiual Achievement Test, Woodcock-Johnson, Passage Comprehension)	0.08	0.43✢	

✢<.10; *p<.05; **p<.01
Note. Effect sizes estimates reported by study authors.

Toward A Systematic Program of Research at Grade 3

Two of the reported studies in Table 1 were conducted by the research team, of which I was a member for the past six years. The studies were conducted as part of a large IES funded collaborative research initiative tasked with developing a comprehensive nonfiction comprehension intervention for struggling readers in the intermediate grades. The initial goal of the research initiative was to develop a single program for students in grades 3-5. However, after three years and five randomized control trials, it became clear that third-grade students needed a different program, which is to say their own program. The results from the studies conducted by this research team corroborated findings from the meta-analysis described above—reading comprehension interventions have been less effective for third-grade students than for students in higher grades.

Developing a separate curriculum for third grade. To improve third-grade students' performance, I created a new set of texts and instructional materials that would be more comprehensive in scope and better aligned to the needs of struggling third-grade readers. Whereas we had primarily been measuring intervention effects with commercially-developed measures in our previous studies, for this study we developed researcher-created measures, more aligned to our intervention to better detect intervention effects across dimensions of instructional transfer, not to supplant the commercially-developed tests but to supplement them (Fuchs et al., 2018). We piloted the new curriculum and assessments with 56 students in 2016 (Walsh, Fuchs, Patton, & Fuchs, 2019).

Figure 1 displays results from the pilot study. The disparate effects on the Gates-MacGinite Reading Test-4, Form S (Gates, MacGinite, MacGinite, Maria, & Dryer, 2012) and the Wechsler Individual Achievement Test-3, Reading Comprehension subtest (WIAT, Wechsler , 2009) highlight the difficulty of measuring the reading comprehension construct and identifying at-risk readers in third grade. At pre, the Gates was used as a screener to determine study inclusion. By design, students were selected because they had norm referenced scores at pretreatment below the 50th percentile on the Gates. The WIAT was also given at pretreatment, but scoring was not completed until two weeks into the intervention period. Norm referenced scores at pre- for many students on the WIAT were above the 50th percentile, and, in some cases, well-above. As Figure 1 illustrates, the differences between the commercial measures was further exacerbated at post-treatment. Results on these two distal, or "far-transfer," commercial tests underscored the need for further work, but also validated our decision to include more proximal researcher-created measures. Based on student performance in the pilot study, we understood the need to continue to try to improve program effectiveness across degrees of instructional transfer.

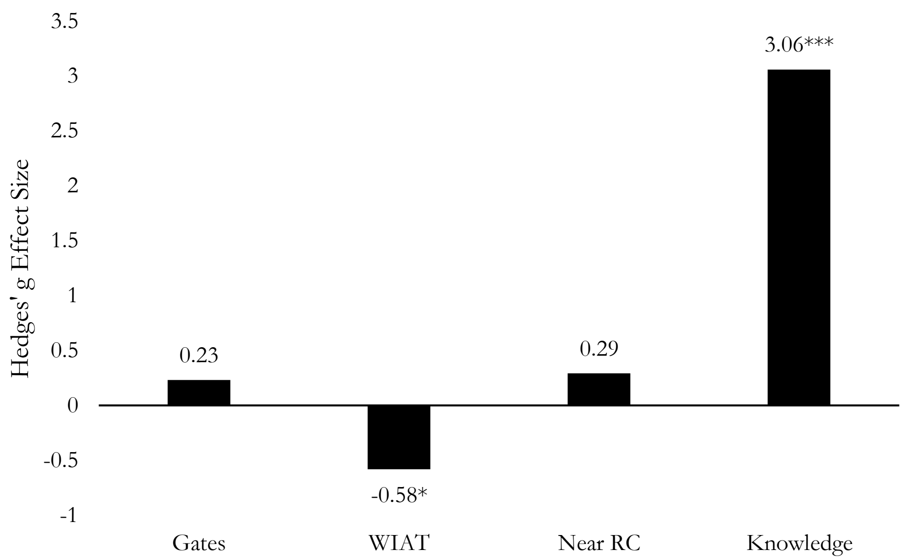


Figure 1. Effect estimates from third grade pilot study (Walsh, 2019).
Gates was the Gates-MacGinite Reading Test-4, Form S (MacGinite, MacGinite, Maria, & Dryer, 2012). WIAT was the Weschler Individual Achievement Test-3, Reading Comprehension subtest (Weschler, 2009). Near RC was a researcher-created reading comprehension measure. Knowledge was a researcher-created knowledge measure. Both the Near RC and Knowledge measures were administered at posttest only. Hedges' *g* effect size estimates calculated using regression coefficients estimated from hierarchical linear models controlling with fixed effect estimates for pretreatment score and random effects for clustering in pairs and classrooms.

Study Purpose

The purpose of this study was to build on the program developed for the pilot study to improve the efficacy and feasibility of an intervention that would improve understanding of informational texts among third-grade at-risk children. The research team and I had four hypotheses: First, by streamlining and improving the instructional activities in our program, we anticipated the intervention would be implemented with fidelity and would yield important effects on measures of student learning acquisition. That is, treatment students would be able to display the knowledge, vocabulary, and strategies specifically taught in intervention on researcher-created measures aligned to treatment. Second, we anticipated that student learning would result in proximal transfer, with treatment students outperforming controls on researcher-developed nonfiction comprehension measures. Third, we hypothesized that the intervention would yield small, but significant effects on far-transfer comprehension measures.

CHAPTER 2

METHODS

Participants

Research staff. All assessments and instruction were administered by the research staff, which included two doctoral students, one fulltime project coordinator, and five graduate research assistants. Both of the doctoral students held master's degrees in Special Education. The project coordinator held a master's degree in education. The graduate research assistants were pursuing master's degrees in Human Development and Counseling (*n*= 3), Leadership, Policy and Organization (*n* = 1), and Special Education (*n* = 1).

Schools. We recruited teachers and students from 14 public elementary schools in a large urban district in the southeastern United States. Most participating schools (86%) were Title 1 schools (at least 40% of students categorized as low-income). Participating schools served students in pre-kindergarten through fourth grade. The average percentage of students achieving English Language Arts proficiency standards during the 2017-2018 school year, across the 14 schools was 27.11% (*SD*= 0.12). Per district policy, all schools were required to use a balanced literacy program for core instruction, to implement an MTSS framework, and to designate an intervention period during the school day.

Teachers. Forty-four third-grade teachers participated in the study. Teachers had between 1 and 6 students from the sample in their classes (equally divided between treatment and control when possible). Teachers were mostly white (68%), females (93%), and a majority (57%) held advanced degrees. Teacher experience ranged from 1 to 40 years (*M*=11.24 years, *SD*=9.64 years) and also varied with respect to teaching at third grade (*M*=3.97 years, *SD*=4.19). All teachers held elementary education licensure. A few held additional endorsements in the following areas: math (*n* = 3), reading (*n* = 5), special education (*n* = 1), English language learning (*n* = 1), curriculum and instruction (*n* = 1), technology (*n* = 2), library science (*n* = 1), and administration (*n* = 2). Two teachers held National Board Certification.

Student selection and assignment. To identify students at risk for comprehension deficits, we used a gated selection procedure. First, participating teachers submitted the names of students with schedule availability and English language proficiency who they believed would benefit from nonfiction comprehension intervention. We received parental consent to screen 209 of these nominated students.

To ensure that participants had adequate word-reading for participation in a comprehension-focused intervention, we selected students with scores above the 25th percentile on the Test of Word-reading Efficiency-2, Sight Word Efficiency subtest (TOWRE; Torgesen, Wagner, & Rashotte, 2012). Those students who met the word-reading criterion (*n*=138) were given a nonfiction comprehension screening measure. This measure included the four expository passages from the Gates, Form T (Gates T) and a researcher-created nonfiction passage. The purpose of the screener was to ensure students had specific difficulty with nonfiction comprehension.

Students (*n*=93) who correctly answered fewer than 14 items (<50%) on the screener were retained in the sample. Once they had been identified, we administered the full Gates, Form S (Gates S) under standardized conditions to obtain a norm-referenced score. This was done to confirm risk status (score <60th percentile). In addition to these measures, we also administered the Weschler Abbreviated Scale of Intelligence-2 (WASI-2; Weschler, 2011). Eight students were eliminated because they failed to achieve a minimum T-score of 36 on both the vocabulary and matrix reasoning subtests. Eighty-five students met the inclusion criteria and were randomly assigned within school and classroom to either the comprehension treatment or control group. After randomization, conditions were checked for pretreatment comparability. There were no significant differences between the two conditions on screening measures or demographics (see Table 2).

Table 2

Student Descriptive Data by Condition

	Tutored		Control	
Variable	*n*	%	*n*	%
Race				
African American	16	41.03%	11	28.21%
Caucasian	7	17.95%	6	15.38%
Hispanic	14	35.90%	14	35.90%
Other	2	5.13%	8	20.51%
Sex				
Female	25	64.10%	24	61.54%
Male	14	35.90%	15	38.46%
Disability				
IEP	1	2.56%	1	2.56%
No IEP	38	97.44%	38	97.44%
Free and Reduced Lunch				
Qualifed for assistance	26	66.67%	21	53.85%
Did not qualify	12	30.77%	18	46.15%
Missing	1	3.45%	0	0.00%
English Language Proficiency				
ELL	5	12.82%	3	7.69%
Not ELL	34	87.18%	36	92.31%
Retention Status				
Retained	0	0.00%	0	0.00%
Not Retained	37	94.87%	38	97.44%
Missing	2	5.13%	1	2.56%
School-based Tiered Intervention (Fall)				
Tier 2	8	21%	12	31%
Tier 3	1	3%	4	10%
No Tiered Intervention	30	76%	23	59%
Screening Measures	*M*	*SD*	*M*	*SD*
TOWRE 2 SWE PR	47.44	15.68	49.82	15.84
Comprehension Screener RS	9.26	2.30	9.72	2.73
Gates PR	29.41	14.04	28.51	12.22
IQ	91.46	11.24	89.15	7.82

Note. TOWRE-2 SWE PR is Test of Word-reading Efficiency-2, Sight Word Efficiency subtest (Torgesen, Wagner, & Rashotte, 2012) percentile. Comprehension Screener is the combined raw score on the 4 expository passages from the Gates-MacGinite Reading Test-4, Form T (MacGinite, MacGinite, Maria, & Dreyer, 2001) and a researcher-created mid transfer measure. Gates PR is the Gates-MacGinite Reading Test-4, Form S percentile rank (MacGinite, MacGinite, Maria & Dreyer, 2001). IQ scores were calculated based on performance on the Weschler Abbreviated Scale of Intelligence (Weschler, 2011).

Student descriptive information. As intended, students' word-reading scores were in the average range based on the TOWRE (Mean Percentile=48.63, *SD*=15.70) Student comprehension was in the low average range. The mean percentile of the sample on the full Gates S was 29.26, and only 7 students had scores above the 50th percentile. The average score on our nonfiction comprehension screening measure was 9.49 (out of a possible 24). Mean T-scores on the WASI-2 were in the low average range for both vocabulary (*M*=45.99, SD=8.20 and matrix reasoning (*M*=43.00, SD=7.78). Based on these subtests the mean IQ of the sample was 90.31 (*SD*=9.69). The study sample was demographically diverse reflecting the diversity of the large urban district from which they were drawn. Only two students had documented educational disabilities. None of the students in the sample had been retained.

Attrition. During the intervention, seven students were lost to attrition (8%). Three students (7%) were lost from treatment and four (9%) from control. Overall and differential attrition rates were below the conservative threshold established by the What Works Clearinghouse (WWC; 2008). Six of these students moved to another school, district, or state. One treatment student qualified for math intervention and was removed from the treatment by their school's response to intervention coordinator because of scheduling conflicts. The final sample included 78 students.

Study Design

The study was a randomized control trial. During pretesting, students were randomly assigned using a stratified procedure (within classroom and school). Treatment students were assigned to pairs based on schedule availability and word-reading ability. Students randomly assigned to the treatment condition received 14 weeks of paired tutoring intervention. Control students were not observed or assessed during the intervention period. After tutoring, all students were post-tested.

Intervention Procedures

All students in the treatment condition were assigned to pairs. Pairs participated in 42 tutoring sessions with a trained member of the research team. Sessions were designed to be 45-50 minutes in length. Students were pulled from their classroom and tutored in another quiet location in their school. Initially, all tutoring was scheduled to be delivered three times per week during teacher-designated intervention blocks, but disruptions due to student illness, school event, and inclement weather occasionally resulted in small schedule changes. Because three treatment students were lost to attrition, their partners were tutored as singletons rather than pairs for a portion of the intervention period.

Materials. Lessons were fully scripted to facilitate implementation fidelity. Daily lesson scripts were displayed in easels with a tutor view on one side and a student view on the other. Strategies were taught with the use of visual cards arranged on a ring (see Appendix A). As students developed strategy facility, tutors gradually faded the visual supports. Other tutor materials used during the lessons included laminated graphic organizers (also Appendix A), game materials, timers, strategy stickers, and dice. Tutors were also provided with electronic media libraries to accompany each text. In addition to the tutor materials, each student received a workbook, binder of accompanying worksheets, and behavior point card which were used during each lesson.

Behavior management. Tutors reviewed the routines and rules for tutoring at the beginning and end of each session. The rules were: Work hard. Be respectful. Be a good team member. Tutors used a point system and were encouraged to generously award points when students followed these rules. Tutors were instructed to provide specific praise when awarding points. Students traded accumulated points for small prizes throughout the intervention period.

Fidelity. Before working with students, tutors received six hours of training, were assigned a practice partner, and required to deliver a lesson to the project coordinator with greater than 90% fidelity. During the intervention period, fidelity was supported and maintained in several ways. First, tutors received follow-up training and shared program related concerns during weekly team meetings. Second, each tutor was observed three times (approximately during weeks 4, 9, and 14) in-situ by the project coordinator and at least once with each student pair. Adherence fidelity was collected during these observations using a 38-item checklist. Tutor quality was also rated during each of these observations. Each tutor was rated as either "highly effective," "somewhat effective," or "ineffective" in lesson pacing, use of correction procedures, and behavior management. In addition to these in-situ observations, tutors audio-recorded and completed a record form for each session. Record forms included the session duration, a lesson completion checklist, and data on student strategy performance.

Reading PI: Instructional Model

We incorporated four evidence-based practices into our instructional model: high quality text, explicit instruction, peer mediation, and scaffolding for situation models. These components were intended to support student acquisition of seven strategies believed to be important for informational text comprehension in third grade. Each of the components and strategies is described below and visually represented in Figure 2.

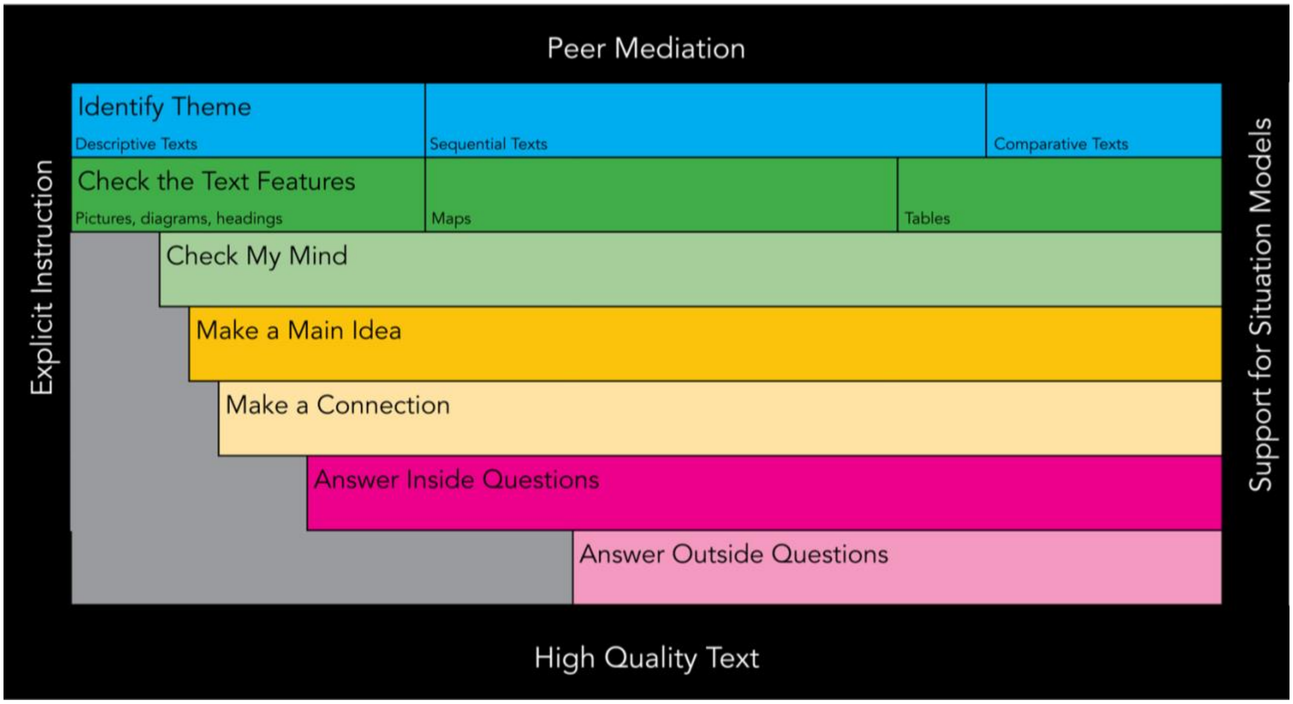


Figure 2. Instructional components included in the third-grade nonfiction reading comprehension intervention.
The components listed on the black frame are the evidence-based practices used to facilitate student acquisition of the comprehension strategies. Strategies taught are shown in the center of the figure. Students were taught to think about the theme or the author's message (blue bar) throughout reading. Green strategies were used before, yellow strategies during, and pink strategies after reading. Strategies in darker shaded bars were meant to help students attend to important aspects of the text. Strategies in lighter shades were intended to help students develop coherent, well-elaborated situation models.

High quality texts. We wrote 19 original, informational texts for the program. Each of which was 2-6 paragraphs. To ensure that texts were written at an appropriate level for struggling third-grade readers, we used several readability metrices: Lexile level (Stenner, Smith, & Burdick, 1983), Flesh-Kincaid grade equivalent (FK, Kincaid, Fishburne, Rogers, & Chissom, 1975), and Coh-metrix readability indices (Graesser, McNamara, Louwerse, & Cai, 2004). Readability statistics for each text are summarized in Appendix B. The average Lexile level of the texts in this intervention was 582L (*SD*=72L, Minimum = 470L, Maximum = 720L). The average FK grade equivalent for the intervention texts was 3.8 (SD = 1.1). Coh-metrix (Graesser, McNamara, Louwerse, & Cai, 2004) provides several additional indices for quantifying readability based on content features such as language and cohesion. In comparison to the wider corpora of possible texts, the texts in this intervention had high syntactic simplicity and word concreteness (mean percentiles were 83 and 70 respectively). The texts had average narrativity, referential cohesion and deep cohesion (mean percentiles were 59, 48, and 44 respectively).

In addition to ensuring that each text was at an appropriate reading ability level, we also worked to ensure that each text provided ample opportunities to practice comprehension strategies, engage in discussion, and build high quality situation models. Appendix C summarizes the instructional components incorporated into each text. Texts were arranged in thematic units designed to systematically build content knowledge: *Your Amazing Brain, Superkids*, and *Exploring Space*. Each unit was presented in a separate student workbook complete with title page, table of contents, and glossary. Workbooks and texts were formatted to resemble authentic third-grade texts (see Appendix D for an example). Common informational text features, such as pictures, headings, diagrams, maps, tables, and charts, were incorporated into each text. Each unit also featured a different informational text structure: descriptive, sequential, or comparative.

Peer mediation. Peer mediation was employed in an attempt to facilitate guided practice and allow students opportunities to engage in authentic discussion. Students took turns being “Coach” and “Reader” throughout the lessons similar to the method described in Fuchs and colleagues (2001). The Reader read or practiced a strategy while the Coach provided feedback and support. To encourage additional peer interaction, tutors were provided with open-ended questions and media discussion prompts designed to promote discussion and engagement. Stickers and points were used to further encourage teamwork and strengthen the peer dynamic.

Scaffolded support for construction of situation models. We designed the intervention to support student construction of situation models in three ways: First, we designed the text units to systematically build students’ conceptual and vocabulary knowledge (Compton et al., 2014). Second, four of the strategies taught in the intervention were specifically meant to help students make coherence or elaborative inferences. Last, we embedded questions and corrective feedback to implicitly guide students in the development of coherent situation models. Despite these supports, we anticipated that many students would have significant, idiosyncratic gaps in background knowledge that would impede their comprehension and attenuate intervention effects. To address this, we curated media libraries to supplement each text.

Explicit strategy instruction. Explicit instruction incorporates directions in clear, direct language; modeling of efficient solution strategies; support for potential gaps in student background knowledge and prerequisite skills; gradual fading of support for correct execution of strategies; opportunities for student practice; and systematic, cumulative review (Fuchs, Fuchs, & Malone, 2017). This program relied on these principles to presented seven comprehension strategies for use before, during, and after reading informational texts. Description of these strategies follows.

Nonfiction Comprehension Strategies

Answer the theme question. In this program, the author's message was referred to as the theme. Students were taught to approach a text as a private investigator (PI) might approach a crime scene, using all of the tools, or strategies, at their disposal to help them solve the case (i.e. identify the theme). In the service of identifying the theme, students were taught to identify and utilize a texts' structure. Throughout the reading process, tutors and students used text structure graphic organizers (Appendix A) and structure-specific guiding questions to facilitate theme identification.

Check the features. Students were taught to use the "Check the Features" strategy before reading to access important information from text features and to make predictions about the author's theme. Tutors and students use the "Check the Features" guide (Appendix A) to support implementation of this strategy. Part of this strategy involved student's self-administration of a vocabulary check-up where they checked their background knowledge, the context, and a glossary to learn vocabulary terms. Tutors used follow-up questions to ensure understanding.

Check my mind. After "Check the Features," students were taught to use the "Check my Mind" strategy to activate and evaluate their background knowledge related to the theme. Students asked each other, "What do you already know about [the predicted theme]?" Then, students evaluated their collective background knowledge. When students recognized that they had inadequate background knowledge, they could choose to access the media library.

Make a main idea. The first strategy used during reading was the "Main Idea" strategy. It had three steps: 1) Name the most important who or what. 2) Tell the most important thing about the who or what. 3) Say the main idea. When students identified the main idea of a paragraph, they wrote it on the graphic organizer that corresponded to the structure of the text. Previously created paragraph level main ideas were reviewed after each paragraph to help reinforce central information from earlier in the text and improve recall of literal information.

Make a connection. Whereas the "Main Idea" strategy was meant to help students focus on important information in the text, the second strategy used during reading, "Make a Connection," was meant to help them develop and elaborate their situation models. During the first unit, tutors facilitated connection-making implicitly. They highlighted pronouns in each paragraph and checked that students made the correct anaphoric connections. They then used a question to check that students had made the deeper connections necessary for building a coherent situation model. For each of these questions, feedback tables included in lesson scripts helped tutors provide appropriate, responsive support to students. In the second unit, students were taught to identify pronouns and anaphora on their own.

Answer questions. After reading, students used the "In and Out" strategy to help them answer both factual and inferential questions about the text. In the first unit, students were taught to first read the questions and look for evidence "inside" the text to answer factual questions. In the second unit, students were taught to go "outside" of the text and use their background knowledge when the evidence provided in the text was insufficient to answer the question.

Control Condition

Students in the control condition were not formally assessed or observed during the intervention period. Instead teacher surveys, administered electronically in the fall and spring, provided a partial picture of the BAU condition. Tutoring was originally scheduled to occur during reading intervention or literacy periods as designated by teachers. However, teachers reported that only a minority of treatment students (n = 12 (30%) in the fall and n = 16 (40%) in the spring) missed literacy instruction by participating in our intervention. One student missed some Tier 3 intervention in the spring. Six students in the fall and nine students in the spring missed core (Tier 1) reading instruction. Only six students in both the fall and spring missed some Tier 2 intervention. Survey responses did not indicate the nature of other activities provided in the business as usual condition in lieu of literacy instruction. Teachers confirmed that no student missed math instruction, lunch, recess, or related arts while participating in our intervention.

Although we conceptualized our reading comprehension program as appropriately intensive for students in Tiers 2 or 3 of an RTI or MTSS framework, most of our sample was not identified by school personnel for Tier 2 intervention. The small number of children placed in Tier 2 for instruction (treatment n = 8, control n = 12) or Tier 3 (treatment n =1, control n=4) generally received a multi-component intervention that included both comprehension instruction and word-reading support. Such instruction was delivered with similar duration and frequency as our treatment, but it was conducted in much larger groups (4-40 students). As reported by teacher surveys, there was no demonstrable difference between the intensity of Tier 2 and Tier 3 in our study schools. A small number of students (4 in the fall and 5 in the spring) received additional tutoring in reading outside of tiered intervention support or our treatment. There were no significant differences between conditions in the number of students who participated in additional tutoring.

Measurement

Test procedures. Students were assessed before and after treatment on a variety of measures (described below). All measures were administered by members of the research team blind to study assignment. Testing occurred in five sessions before and five sessions after treatment. Each session took between 10 and 45 minutes to complete. Students were assessed either individually or in small groups (up to four students). Prior to administration in the fall and spring, research assistants were required to deliver and score each measure with a minimum of 90% fidelity. All test protocols were double-scored and double-entered by different members of the research team. All testing sessions were audio-recorded. Scoring reliability was completed from audio files for 20% of all test sessions. Reliability information and pre/post measure correlations for all measures are provided in Table 3.

Table 3

Measure Reliability Statistics

Measure	Scoring Reliability		Sample-based Cronbach's alpha		Pre/Post Correlation		Publisher reported Reliability
	Pre %	Post %	Pre α	Post α	Treat r	Control r	
Reading Comprehension Measures							
Knowledge			0.05	0.66	-0.01	0.20	
Main Idea	0.94	0.91	0.68	0.80	0.17	0.40	
Near RC			0.28	0.58	0.23	0.38	
Mid RC			0.22	0.50	0.20	0.37	
Gates T			0.19	0.57	0.33	0.29	K-R 20 = 0.94
Gates S			0.64	0.76	0.67	0.54	K-R 20 = 0.93
Other Measures							
TOWRE	0.99	0.98			0.54	0.77	α = 0.90
Vocabulary	0.87		0.71				α = 0.87, 0.91
Matrix Reasoning	1.00		0.48				α = 0.87, 0.91

Note. Scoring reliability was calculated as (agreements /(agreements + disagreements)) for 20% of all assessments. Scoring reliability was not calculated for multiple-choice measures. Knowledge is the researcher-created knowledge and vocabulary measure, Main Idea is the researcher-created main idea generation measure, Near RC is the researcher-created near Transfer reading comprehension measure, Mid RC is the researcher-created mid transfer reading comprehension measure. The Gates T refers to the 4 expository passages taken from the Gates MacGinite Reading Test-4, Form T (MacGinite, MacGinite, Maria, & Dreyer, 2001). Gates S is the full Gates Macginite Reading Test-4, Form S (MacGinite, MacGinite, Maria, & Dreyer, 2001). TOWRE-2 SWE is Test of Word Reading Efficiency-2, Sight Word Efficiency subtest (Torgesen, Wagner, & Rashotte, 2012). Vocab and Matrix were the two subtest of the Weschler Abbreviated Scale of Intelligence-2 (Weschler, 2011). IQ measures were only given at pre-treatment. K-R 20 is the Kuder-Richardson Formula 20.

Acquisition. Two measures were developed by the research team to assess learning acquisition. They were administered at pre- and posttreatment. The first test, Knowledge and Vocabulary (Knowledge), required students to answer 15 multiple-choice questions about concepts and vocabulary presented in the treatment instructional passages. The tester read aloud questions and answer choices, proceeding one question at a time so that all students could mark an answer in their test booklets. This test was administered in small groups. The second acquisition measure was Main Idea Generation (Main Idea) and required students to read two passages aloud, each consisting of four paragraphs. After each paragraph, students were asked to state a main idea. This test was administered individually.

Near- and mid-transfer reading comprehension measures. Two researcher-created, proximal measures were used to assess students' abilities to transfer knowledge and strategies learned in tutoring to comprehend similar but unfamiliar informational passages. Both tests were administered in small groups. The Near-Transfer Reading Comprehension (Near RC) was researcher-developed and required students to read three passages and answer a series of multiple-choice questions. The passages and questions were similar in presentation (e.g., layout and design) to those used during treatment. The test passages had *not* been seen previously by the students, but their content was drawn from topics that had been addressed in tutoring. The Mid-Transfer Reading Comprehension test (Mid RC) was similar to the Near RC measure in several respects. Like the Near RC measure, students read a single passage and answered written multiple-choice questions. Unlike the Near RC measure, the topic and text structure of the passages were unfamiliar to students. The Mid RC measure was used at pretreatment as part of the screening measure.

Far-transfer measures. Two versions of a commercial measure, The Gates-MacGinite Reading Test-4 (MacGinite et. al, 2000), were used to assess far transfer reading comprehension growth. Both were administered in small groups. The Gates T was a selection of 4 non-fiction passages and 16 multiple-choice questions excerpted from Form T of the Gates. This measure was also used at pretreatment as part of the screening measure. The standardized administration procedures for the full test call for time limits, however time limits were not used for the Gates T administration. We also administered the full Form S of the Gates. On the Gates S, students read 11 short passages and provided written answers to multiple-choice questions. Per standard administration procedures, students were given 35 minutes for the test. The full Gates S differed from the Gates T because it included fictional passages in addition to nonfiction items and was administered in standard fashion.

Word-reading. Word-reading was assessed using the TOWRE. The TOWRE requires students to read as many sight words as possible in 45 seconds from a list of words that gradually increase in difficulty. ORF requires students to read passages of connected text (typically 2) as quickly as possible. This test was administered individually following the standardized administration procedures. Sample-based Cronbach's alpha is not reported for this measure because it is not appropriate for speeded tests.

IQ. We used two subtests from the WASI-2, Matrix Reasoning and Vocabulary. Both of these subtests were administered individually. Matrix Reasoning assesses nonverbal reasoning with pattern completion, classification, analogy, and serial reasoning tasks. For each item, students select one of five options that best completes a visual pattern. The Vocabulary subtest evaluates expressive vocabulary, verbal knowledge, and foundational information. For each item, students identify a picture or provide a definition for a word read aloud by the tester.

Data Analysis

Method of analysis. Students in the sample were nested and cross classified. That is, all students were nested within schools and classrooms, but students in the treatment arm were also nested in pairs. To deal with inter-dependency arising from this data structure, I used a cross-classified multilevel model suggested by Sterba (2017) for all analyses. I ran a series of unconditional models for each outcome that included school and classroom random effects for the full sample, and pair random effects for the treatment condition. ICC results from the unconditional models for each outcome are provided as Appendix F.

Final models for each outcome had the following covariates: pretreatment score on the corresponding outcome measure, pretreatment TOWRE score (to account for the method of stratified assignment), a dummy variable comparing treatment students to control students, and any necessary random effects as determined by nonzero ICCs. Level-1 residuals were estimated separately for treatment and control groups to account for possible heterogeneity across conditions. Standard errors were adjusted by applying the Kenward-Roger adjustment as recommended by Baldwin, Bauer, Stice, and Rhode (2011) to account for the small number of clusters. Models were estimated using the *mixed* command in Stata/Se v.15. Effect size estimates were calculated using Hedges *g* small-sample bias correction, the formula provided by the What Works Clearinghouse (WWC, 2008), and coefficient estimates of the fixed effect comparing treatment students to controls produced by the multilevel models.

Primary analysis. I estimated a series of models, as described above, for each outcome across degrees of instructional transfer (i.e., near-to-far). I first estimated models for acquisition outcomes (Knowledge and Main Idea); then for the researcher-created proximal (near- and mid-) transfer reading comprehension outcomes (Near RC and Mid RC). Finally, I estimated models for commercially-developed far-transfer reading comprehension outcomes (Gates T and Gates S).

Supplemental analysis. To better understand intervention effects, I completed item type analysis. Items on the Near RC measure could be easily categorized as main idea, theme, factual or inference. Thus, separate models applying the methods described above were estimated for each item type on the Near RC measure.

CHAPTER 3

RESULTS

Histograms, scatter plots, summary statistics, and the results of Wilk's tests reveal that most measures were normally distributed at pre- and post-treatment with a few notable exceptions: Post-treatment distributions for Knowledge and Main Idea measures were non-normal and bimodal as might be expected from a very strong treatment effect. There was some evidence of floor effects on the Main Idea measure. The distribution of scores on the TOWRE measure was slightly skewed, with the majority of scores clustered near the study inclusion cut point. Means and standard deviations for all outcome variables by treatment are in Table 4. Correlations among measures are in Appendix E.

Table 4

Pre-/Posttreatment Means on Outcome Measures by Condition

	Treatment (n=39)		Control (n=39)	
Measure	Mean	SD	Mean	SD
Pr Knowledge	5.49	1.75	5.18	1.67
Po Knowledge	10.49	1.99	5.97	1.42
Pr Main Idea	2.00	1.99	1.56	1.62
Po Main Idea	3.67	2.38	1.38	1.66
Pr Near RC	8.08	2.42	8.26	2.83
Po Near RC	11.10	3.43	9.85	3.53
Pr Gates T	6.03	2.07	6.79	1.98
Po Gates T	8.79	2.80	7.87	2.67
Pr Mid Transfer RC	3.23	1.46	2.92	1.49
Po Mid Transfer RC	4.18	1.55	3.87	2.08
Pr Gates S NCE	37.67	8.83	37.18	8.08
Po Gates SNCE	37.54	11.41	38.10	10.33

Note. Knowledge is a researcher-created measure that assess the vocabulary and concepts taught during intervention. Main Idea is a researcher-created measure that tests students' ability to create Main Idea statements for paragraphs. Near RC is a researcher-created near transfer test of reading comprehension. Gates T is the raw score on the 4 expository passages of the Gates-MacGinite Reading Test-4 - Comprehension, Form t (MacGinite, MacGinite, Maria, & Dreyer, 2000). Mid Transfer RC is a researcher-created measure of Reading Comprehension that includes content not taught in intervention. Gates S NCE is the normal curve equivalent score on the Gates-MacGinite REading Test-4 Comprehension, full form S.

Fidelity

Adherence and quality. Data collected during in-situ observations indicate the program was delivered by all tutors with a high level of fidelity across time-points and tutors (mean aggregated adherence = 96.47%, *SD* = 3.93%). For each tutor, average fidelity of implementation was 97%, 99%, 93%, 96%, and 97%. Time-point averages were 94.6% at week 4, 97.0% at week 9, and 97.8% at week 14.

Tutoring fidelity was evaluated qualitatively as well as quantitatively. High quality ratings indicate that tutors delivered the lessons smoothly with few pauses and that they provided appropriate feedback to students on a consistent basis. All tutors were rated as "highly effective" or "somewhat effective" for each of the quality indicators (lesson pacing, use of correction procedures, behavior management) at all time-points. No tutors were rated as ineffective on any indicator. Moreover, all tutors were rated as "highly effective" for behavior management at all time-points. This suggests that tutors not only applied behavior management techniques, but that student behavior and engagement did not compromise instruction.

Dosage. Mean duration of tutoring across all sessions was 46.21 minutes (SD=1.15 minutes). Only one lesson (lesson 16) had an average session duration less than 44 minutes (42.82 minutes). None of the lessons had an average duration greater than 50 minutes. Only one student (tutored individually after their partner was lost to attrition) had an average session duration less than 44 minutes (mean session length=43.37). Average session duration was similar across the three units of instruction (46.28, 46.01, and 46.40). Based on the number of sessions completed (42 for all students) and average session duration, treatment students received approximately 32.35 hours of instruction.

Strategy learning during tutoring. Data taken from tutor record forms indicate variability in the extent to which students demonstrated independent use of strategies during tutoring. Students averaged 83% on the "make a connection strategy" and all students displayed this strategy at least once. Students averaged 72% on the "check the features" strategy, 72% on "making main ideas," 71% on the "check my mind" strategy and 71% on "answering theme questions." Some students did not independently demonstrate any of these strategies. Tutors reported a much lower average percentage for answering "other types of questions" (32%). The maximum average percentage reported for answering questions was only 50%.

Primary Analyses

Acquisition measures. Statistically significant and large effects were observed for treatment on the two acquisition measures, Knowledge and Main Idea. Model results for both measures are provided in Table 5. The treatment effect on content acquisition as measured by the Knowledge measure was g = 2.64 (SE = 0.31, $p < 0.001$). Pretreatment knowledge was not a significant predictor of the post-treatment knowledge score. The effect on main idea was similarly significant and large (g = 1.00; SE = 0.24, $p < 0.001$). Unlike on the knowledge measure, however, pretreatment score was a significant covariate in the Main Idea model ($p < 0.01$).

Near- and mid-transfer. Treatment effects on the near- and mid-transfer measures were mixed. On the Near RC measure, the effect was marginally significant and moderately large (g = 0.37; SE = 0.23; p =0.10). There was no reliable effect on the Mid RC measure (g = 0.03; SE = 0.23; p = ns). Pretreatment scores significantly predicted outcomes on both measures (see model results for Near RC and Mid RC in Table 5).

Table 5

Multilevel Model Results for Post-treatment Researcher-Created Acquisition and Proximal Transfer Measures

	Acquisition Measures								Proximal Transfer Measures							
	Knowledge				Main Idea				Near RC				Mid RC			
Effects	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*
Fixed Effect																
Intercept	4.02	2.90	1.38	0.17	0.60	3.83	0.16	0.88	9.14	6.42	1.42	0.16	2.66	3.09	0.86	0.39
Pretreatment Score	0.11	0.11	1.07	0.29	0.33	0.13	2.63	0.01	0.42	0.15	2.70	0.01	0.34	0.13	2.66	0.01
Pretreatment TOWRE	0.01	0.03	0.47	0.64	0.00	0.04	0.12	0.91	-0.03	0.07	-0.43	0.67	0.00	0.03	0.12	0.90
TRT vs Control	4.61	0.37	12.39	0.00	2.34	0.51	4.57	0.00	1.30	0.77	1.69	0.10	0.06	0.37	0.17	0.87
Random Effect																
Teacher (τ_u)	0.40				0.71				0.00				0.95			
Pair (τ_r)	0.00				0.00				0.00				0.00			
Residual (treatment) (σ_{12})	3.46				6.59				11.79				1.45			
Residual (control) (σ_{22})	1.73				2.30				10.64				2.74			

Note. (n = 78) for all models. Knowledge was the researcher-created measure of content acquisition. Main Idea was the researcher-created measure of main idea strategy acquisition. Near RC was a researcher-created measure of reading comprehension on texts that were similar in content and structure to texts read in intervention. Mid Transfer was a researcher-created measure of reading comprehension with unfamiliar content. TOWRE was the Test of Word-reading Efficiency-2, Sight Word Efficiency standard score (Torgeson, Wagner, & Rashotte, 2012). Table represents results for cross-classified hierarchical linear models with random effects for teacher and pair. For all models *n*=78.

Table 6

Multilevel Model Results Posttreatment Commercial, Far Transfer Measures

	Far Transfer Measures							
	Gates T				**Gates S**			
Effects	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*
Fixed Effect								
Intercept	0.38	4.94	0.08	0.94	8.85	9.97	0.89	0.38
Pretreatment Score	0.46	0.14	3.24	0.00	0.76	0.12	6.20	0.00
Pretreatment TOWRE	0.04	0.05	0.89	0.37	0.01	0.10	0.05	0.96
TRT vs Control	1.18	0.67	1.75	0.08	-0.44	1.18	-0.37	0.71
Random Effect								
Teacher (τ_u)	0.00				0.00			
Pair (τ_r)	2.72				0.00			
Residual (treatment) (σ12)	4.06				23.23			
Residual (control) (σ22)	6.88				28.78			

Note. (*n* = 78) for all models. Gates T was the raw score on the expository passages of the Gates-Macginite Reading Test-4, Form T; Gates Full Form S was the standard administration of the Gates-MacGinite Reading Test-4, Form S both (MacGinite, MacGinite, Maria, & Dreyer, 2001). TOWRE SS was the Test of Word-reading Efficiency-2, Sight Word Efficiency standard score. Models estimated using cross-classified hierarchical models with random effects for school and pair. For all models *n*=78.

Far transfer. Both far transfer outcomes were derived from the same commercially-developed measure, yet they yielded different results (see Table 6). Whereas the treatment effect on the Gates T was marginally significant and moderately large ($g = 0.43$; $SE = 0.23$; $p < 0.10$), there was no significant effect on the Gates S ($g = -0.07$; $SE = 0.23$; p = ns). Pretreatment scores were significant covariates for both measures.

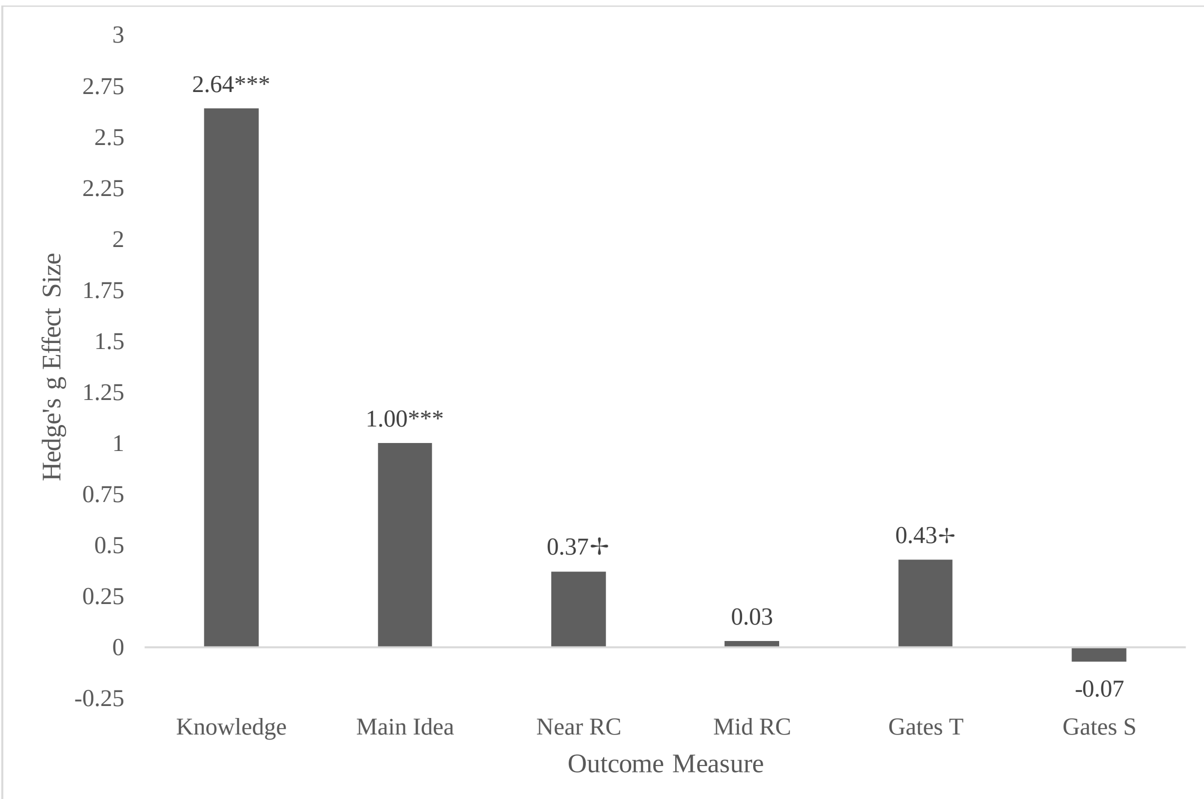


Figure 3. Hedges' *g* effect size estimate (treatment versus controls) by outcome measure.
Effect sizes calculated using the small sample bias adjustment suggested by Hedges' *g*, the formula suggested by the What Works Clearinghouse (2008), and coefficient estimate of the fixed effect for treatment.
**** p<.001; ✢ p<.10*

Supplemental Analyses

Item type analysis suggested that treatment students' performance on the inference items may have contributed to the marginally significant effect on the broader Near RC measure ($g = 0.59$, $p < 0.05$). As displayed in Figure 4, treatment effects on other item types were non-significant: $g = 0.00$ on main idea questions; g = 0.25 on theme questions, and g= 0.04 on factual questions. The model results for each item type are in Appendix G.

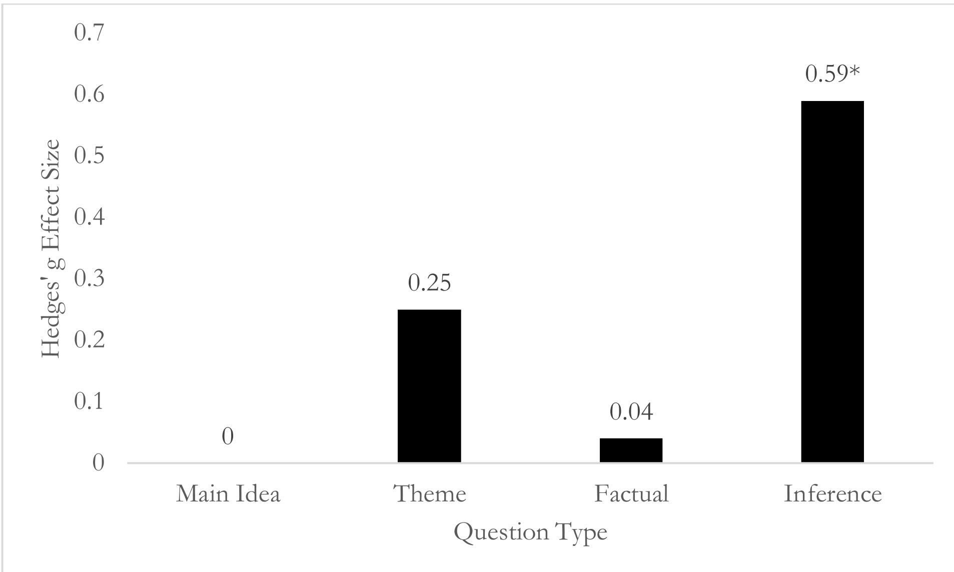


Figure 4. Hedges' *g* effect size estimates (treatment versus controls) by item type.
Item types from the researcher-created Near RC measure. Effect sizes calculated using the small sample bias adjustment suggested by Hedges' *g*, the formula suggested by the What Works Clearinghouse (2008), and coefficient estimate of the fixed effect for treatment.
* $p<.05$

CHAPTER 4

DISCUSSION

The purpose of this study was to evaluate the efficacy of a third-grade reading comprehension intervention focused on informational text. We tested the efficacy of the program by means of a randomized control trial involving 78 third-grade students drawn from 44 classrooms in 14 public elementary schools in Nashville. Students were chosen because they were at-risk for developing comprehension deficits despite having low-adequate word-reading skills. Students in the treatment condition were assigned to pairs and they participated in 42 sessions with trained tutors. The tutoring program used explicit instruction, peer mediation, and high-quality texts to teach seven strategies we believed would support the children's comprehension of informational text. Multiple outcome measures were used to evaluate program efficacy and transfer of learning to familiar-to-novel contexts.

Was the Reading Comprehension Intervention Effective?

As expected, large and statistically significant effects were obtained on measures of learning acquisition ($g = 2.64; g = 1.00$), which indicated that treated students learned and retained content and strategies taught during the tutoring sessions. Results on the Near RC measure was marginally significant ($g = 0.37$, $p = 0.10$), suggesting treated students improved their comprehension of informational texts more than controls on passages resembling but not identical to passages used for instruction. There was no treatment effect on the researcher-created Mid RC measure. Student performance on the two Gates far-transfer measures were inconsistent: marginally significant effects on the (expository test items) of the Gates T ($g = 0.43$, $p = 0.08$) favoring treated children; no treatment effects on the Gates S.

This study was our fifth iteration of a comprehension intervention for struggling readers in the intermediate grades (grades 3-5), and the sixth randomized control trial in which our research team included third-grade students (see Fuchs et al., 2018; Walsh et al., 2017; Walsh et al., 2019). Previous efforts yielded few treatment effects at grade 3, despite promising findings for participants in grades 4-5. Viewed against a backdrop of this prior work, findings from this study may be seen as a step (however modest) in the right direction; that is, moving closer to an effective third grade reading comprehension program for informational texts.

We recognize, however, that many might disagree with this interpretation, and disagree for a specific reason. The *zeitgeist* in the research community has long favored commercially-developed measures over researcher-created tests when evaluating program efficacy (Clemens & Fuchs, 2020). Our own Knowledge, Near RC, and Main Idea measures would be described, no doubt, by many researchers and policymakers as "overly aligned" with our intervention. As such, they would be judged invalid indices of treatment effects. They might even be characterized as fidelity-of-treatment measures rather than outcome measures (e.g. Slavin, 2019). Indeed, the Gates T (comprised of expository items) would probably be discounted as overly aligned, too, and the absence of effects on the unaligned Gates S would be taken as evidence of the program's ineffectiveness. Such an interpretation would be a mistake for two reasons.

The Difficulty of Measuring Reading Comprehension Growth

First, most agree that reading comprehension is a complex construct involving many cognitive processes and that it is very difficult to measure. Commercial tests of comprehension can assess different aspects of the construct with varying degrees of seriousness or comprehensiveness. In a recently completed pilot study (Walsh et al., 2019), a significant interaction was obtained between student attention and treatment such that, on the Gates but *not* on the WIAT, there was a 9 point difference in performance between control students with good and poor attention, whereas there was no such difference among the treated children. The Gates, it seems, was more sensitive to student differences related to attention. Therefore, scores on Gates might be assessing student ability to attend as well as their ability to answer multiple choice comprehension questions. A similar moderation effect was found in other studies (see Hendricks, 2020; Walsh et al., 2017).

Consistently low correlations between commercially-developed measures in this study and many others studies further show that commercially-developed measures represent different aspects of the comprehension construct (e.g. Fuchs et al., 2018, Keenan & Meenan, 2014). This may be especially true for young third-grade readers whose comprehension deficits are not attributable to word-reading alone (Keenan & Meenan, 2014).

Regardless of whether we rely on commercially-developed far transfer tests or researcher-made near transfer measures, we can only measure artifacts of reading comprehension (Fletcher, 2006; Clemens & Fuchs, 2020). That is, unlike word-reading which can be measured directly, measuring comprehension virtually always requires inference-making. Such inference-making might lead to overly cautious and pessimistic interpretations of findings when they are based solely on broad, treatment-insensitive, commercial measures (Fletcher, 2006; Clemens & Fuchs, 2020).

The intervention in this study was specifically designed to strengthen students' comprehension of informational texts. As such, a significant portion of the intervention addressed inference-making and the development of coherent situation models for complex informational texts. Higher-order cognitive skills such as inference-making have been found to contribute more to the comprehension of informational text and to responding correctly to inferential question types (Eason, Goldberg, Young, Geist, & Cutting, 2012). The full Gates is designed to measure overall comprehension of narrative *and* informational texts. It includes a mix of literal and inferential question types. It is conceivable that improved performance on Gates items assessing understanding of expository text may have been overshadowed by a lack of improvement on items exploring comprehension of narrative texts on the full measure. This may partially explain the differential findings between the Gates T and Gates S (g = 0.43 and g = -0.07 respectively). The full Gates S, despite its widespread use in previous investigations may be inappropriate for evaluating the importance of a treatment's inference-making component particularly in the context of attempting to strengthen understanding of informational texts.

It would seem that the inclusion of researcher-created proximal measures allows for a more nuanced view of program efficacy. Item analysis on the researcher-created Near RC measure indicated a strong treatment effect on inference items, but not on other question types. This suggests how the program might be strengthened in the future. For example, for subsequent iterations of the program, developers might consider simplifying instruction by deemphasizing less effective strategies (i.e. main idea, answering literal questions) in favor of those that are meant to help children make connections and answer inferential questions, which may have contributed to the improved performance on inference-making. Conversely, in subsequent program development, we might change the instruction of seemingly less effective strategies in hopes of making them more effective. Thus, researcher-created tests like the Near RC measure provide important information to ongoing program development because they measure the specific skills targeted by the program (Catts & Kamhi, 2017; Clemens & Fuchs, 2020). It would be a mistake to ignore promising program effects on researcher-created measures by attending exclusively to treated children's performance on unaligned, commercial measures.

In theory, well-designed researcher-created measures would also allow for more sophisticated analysis (e.g. moderator and mediator analysis; Clemens & Fuchs, 2020). Our sample was too small to support such an analysis due to low power and issues surrounding multiple significance testing. Moreover, our researcher-created measures were imperfect (see Table 3). Preliminary moderator and mediator analysis (conducted but unreported in this paper) proved inconclusive, but they might be an important approach for future research, with the continued development of near- and mid-transfer tests, and if they were conducted with larger samples.

The Difficulty of Remediating Reading Comprehension Deficits at Third Grade

As problematic as it is to measure reading comprehension, it is equally difficult to remediate it, especially in relatively young third-grade students. Given this, even small intervention effects might be functionally important. Recall that the aggregate effect in previously conducted third-grade comprehension interventions was small but statistically significant *g*=0.16 (Walsh, 2020). The effect for all third-grade studies conducted after 2000 was only (*g*=0.06). Effects in the current study on researcher-created measures were much stronger than those in other studies conducted in the last two decades, while the effect on the full Gates S was similar.

Previous research suggests that third-grade readers' developmental immaturity may contribute to the weaker effects obtained for comprehension interventions at grade 3. Third graders have been shown to be less sensitive to inconsistencies in text (McMaster, Espin, & van den Broek, 2014) and less likely apply a known strategy in context (Cross & Paris, 1988). Such performance might partially explain the difference in treatment effects on the Main Idea acquisition measure versus the main idea items on the Near RC measure. Additionally, many third-grade students are still developing foundational word-reading and language skills. Adequate word-reading as determined by norm-referenced measures might not mean that struggling students are ready for an intervention focused exclusively on comprehension, especially if participation means that they do not participate in word-reading instruction. In fact, teachers in this study indicated on surveys that this was the case for a minority of treated students. In our series of studies, we have found considerable evidence to support these developmental differences. (See Table 1 for a comparison of treatment effects between third, fourth, and fifth graders from two of these studies.)

Limitations

There are at least several important study limitations. Several measurement-related issues have already been discussed. There is at least one more. I was unable to estimate inter-rater agreement for fidelity observations. The data needed for this calculation were inadvertently left in the project office during Covid-19 Quarantine. It will be calculated and added to the report at a later date.

There are also sample-related limitations. First, the size of our sample was small particularly at levels 2 and 3 in multi-level modeling (i.e. too few schools and classrooms to fully support multilevel analysis). We were constrained by resources (5 tutors for 14 schools) and a desire to adequately account for random error variation associated with nested data. Our choice to conduct multilevel analysis with such a small sample was conservative and it precluded exploration of moderation and mediation effects. It also may have prevented us from finding statistically significant effects.

Sample selection also had implications for the generalizability of results. We had difficulty identifying a sufficient number of students who met our inclusion criteria (adequate word-reading and deficits in comprehension). We used teacher nominations to help in this regard. However, by utilizing teacher nomination, we may have missed eligible students because teachers were not always aware of their student's comprehension deficits early in the school year. Moreover, school and district policy prevented us from working with most Tier 3 students and others with severe learning disabilities. To meet sample size requirements, our definition of risk was necessarily broad. Thus, we were likely to include students who were likely not at risk, and to exclude others who might have benefited from the intervention. Future investigations should consider targeting students with more intense learning needs.

Another study limitation: We did not observe control students during the intervention period, nor did we observe treatment students participating in other literacy instruction. Teachers tended to provide vague descriptions of their instruction, and this vagueness extended to components of their interventions. This renders a full understanding of the comparison condition impossible. We do know that interventions in which control children participated were often disrupted, because these disruptions also interrupted our tutoring sessions. Disruptions occurred because of frequent school closures due to inclement weather, school assemblies that occurred during the intervention period, and frequent changes to student schedules. We also note that teacher reported group size in their Tier 2 and Tier 3 intervention groups were often considerably larger than the 2:1 ratio in our intervention. This poses an additional limitation related to generalizability.

Finally, this study represents the culmination of a multi-year effort to develop a program designed to strengthen third-grade children's understanding of informational text. In this final effort, we incorporated many evidence-based practices described in prior reviews of comprehension instruction (e.g. Shanahan et al., NRP, 2000, Gersten et al., 2001). Despite the mixed results from the evaluation described in this paper, practitioners might consider using this intervention as a starting point for their instruction. Not because this program is all that it should be, but because there are currently no validated reading comprehension programs for struggling third-grade readers (Walsh, 2020). As a consequence, it is unlikely that many teachers will create more effective programs. The mixed success of our effort to develop a more efficacious intervention indirectly reflects that there is still much to learn about who third-grade children are as learners and how to help those of them who need our help.

.

Appendix A

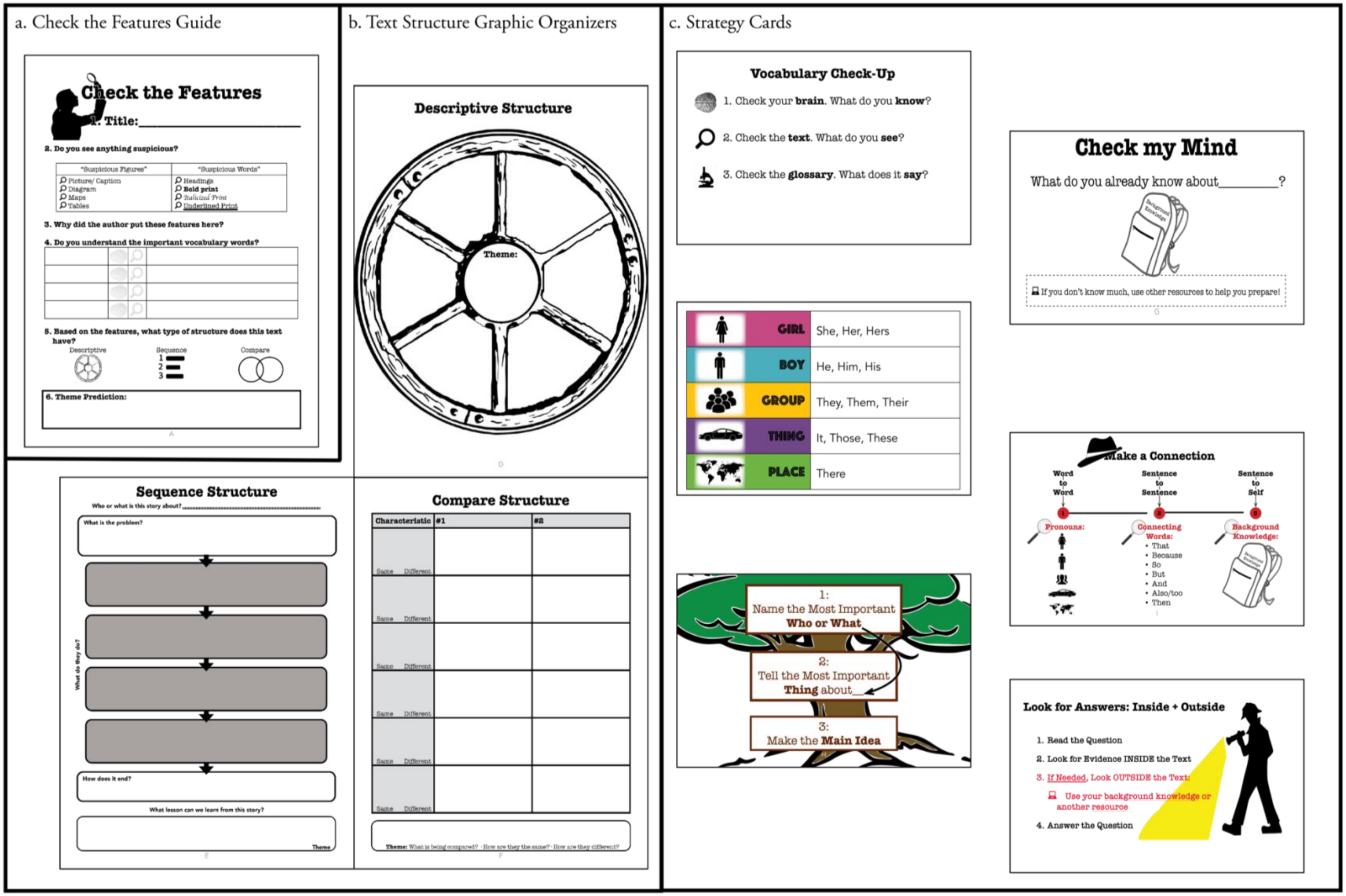


Figure A. Supplemental instructional materials.
Graphic organizers are displayed left. Strategy cards are displayed right.

Appendix B

Table B

Text Readability Statistics

Text	Word Count	Lexile Level	F-K Grade	Coh-Metrix Readability Indices: Narrativity	Syntactic Simplicity	Word Concreteness	Referential Cohesion	Deep Cohesion
Hello, World!	422	490	2.2	70	91	97	99	18
Your Amazing Brain	300	560	3.1	54	67	73	96	13
Take Care of Your Brain	380	500	2.1	89	61	80	99	40
Super Girl: Jessica Watson	245	470	2.8	95	96	67	51	95
Bat Boy: Ben Underwood	189	510	1.7	78	91	58	41	27
Learning to See in the Dark	241	530	2.6	79	87	42	54	15
Meet the Real Spider Girl	203	620	4.6	57	81	66	27	32
Malavath on Top of the World	319	660	4.5	72	82	61	38	39
DJ Focus: Kelvin Doe	198	560	3.6	87	80	83	54	83
DJ Focus Inspires Others	215	650	4.8	71	95	21	20	96
Our Neighborhood	317	590	4.7	17	67	57	83	27
Race to the Moon	355	610	5.0	48	88	68	50	9
Earth Verses Venus	293	650	4.2	24	90	67	56	42
How Does Our Sun Compare to Other Stars?	329	550	3.6	50	62	91	84	17
Life in Space	352	560	4.5	55	96	81	41	47
Trailblazers: Mae Jemison and Guy Bluford	225	690	4.9	41	96	92	65	93
Ham the Astro-chimp	298	720	4.8	37	94	71	41	90
The International Space Station	201	560	4.9	39	81	86	54	15

Note. Lexile Level is a measure of text complexity based on sentence length. F-K Grade is the Flesch-Kincaid Grade Equivalent and is based on mean word length. Higher Coh-metrix readability indices indicate that a text is easier to comprehend. Narrativity is a measure of the degree of narrative components present in a text. Syntactic Simplicity is a measure of the general complexity of sentences. Word Concreteness is a measure of the percentage of concrete, meaningful words in a text. Referential Cohesion is a measure of the degree to which words and ideas are repeated across sentences and paragraphs in text. Deep Cohesion is a measure of the degree to which connectives and causal phrases are utilized throughout a text.

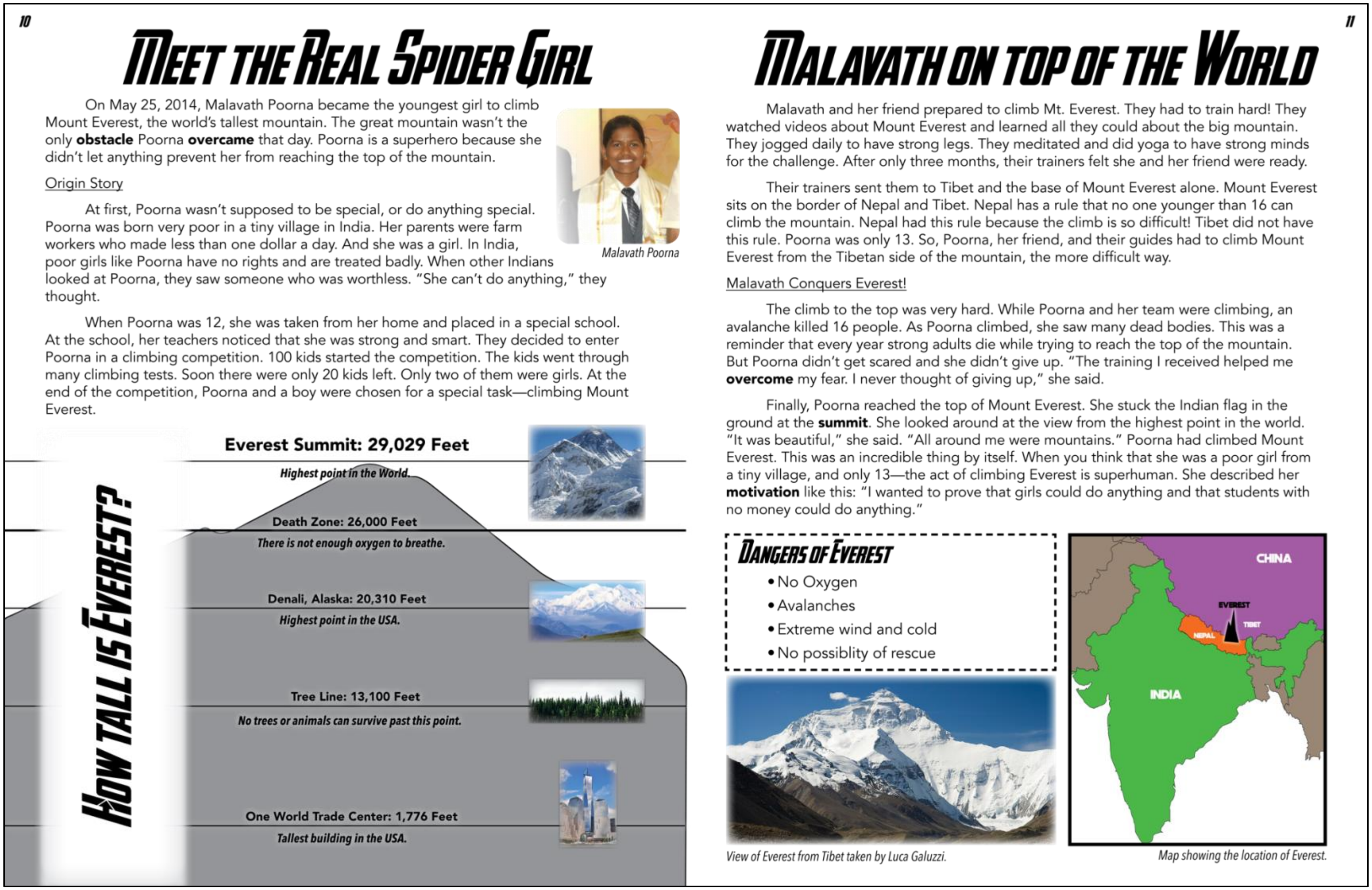
Appendix C

Table C

Instructional Components by Text

Text	Structure	Sessions	Paragraphs	Factual Questions	Inference Questions	Vocabulary Words
Hello, World!	D	6	6			Organ, Nerve, Vibration, Molecule
Your Amazing Brain	D	3	4	11	2	Cerebrum, Cerebellum, Nervous System, Spinal Cord
Take Care of Your Brain	D	3	6	10	2	Cells, Damaged, Chemicals
Super Girl: Jessica Watson	S	3	4	9	2	Inspire, Accomplish, Solo
Bat Boy: Ben Underwood	S	2	3	4	5	Blind
Learning to See in the Dark	S	2	3	4	4	Echolocation
Meet the Real Spider Girl	S	2	3	3	1	Obstacle, Overcame
Malavath on Top of the World	S	2	4	5	4	Summit, Motivation
DJ Focus: Kelvin Doe	S	2	3	2	2	Electric
DJ Focus Inspires Others	S	2	3	3	5	Inventor, Role Model
Our Neighborhood	D	2	3	5	5	Solar System, Star, Planet, Orbit, Asteroid
Race to the Moon	S	2	4	3	5	Launch, Satellite, Rocket
Earth Verses Venus	C	3	6	5	3	Geology, Atmosphere
How Does Our Sun Compare to Other Stars?	C	2	5	3	2	Galaxy, Core, Surface, Nebulae
Life in Space	C	2	5	4	2	Astronaut, Gravity
Trailblazers: Mae Jemison and Guy Bluford	C	1	4	2	3	Shuttle, Engineer
Ham the Astro-chimp	S	1	4	1	3	Chimpanzee, Capsule
The International Space Station	D	1	4	2	3	Rate, Research

Note. Structure refers to text structure; D=Descriptive; S=Sequential; C=Comparative. For each paragraph read, students identified a Main Idea. For the first 5 texts, the tutor facilitated connection making implicitly. Beginning with the 6th text, students began making connections for each paragraph on their own using the Making a Connection strategy.



10

Meet the Real Spider Girl

On May 25, 2014, Malavath Poorna became the youngest girl to climb Mount Everest, the world's tallest mountain. The great mountain wasn't the only **obstacle** Poorna **overcame** that day. Poorna is a superhero because she didn't let anything prevent her from reaching the top of the mountain.

Malavath Poorna

Origin Story

At first, Poorna wasn't supposed to be special, or do anything special. Poorna was born very poor in a tiny village in India. Her parents were farm workers who made less than one dollar a day. And she was a girl. In India, poor girls like Poorna have no rights and are treated badly. When other Indians looked at Poorna, they saw someone who was worthless. "She can't do anything," they thought.

When Poorna was 12, she was taken from her home and placed in a special school. At the school, her teachers noticed that she was strong and smart. They decided to enter Poorna in a climbing competition. 100 kids started the competition. The kids went through many climbing tests. Soon there were only 20 kids left. Only two of them were girls. At the end of the competition, Poorna and a boy were chosen for a special task—climbing Mount Everest.

11

Malavath on Top of the World

Malavath and her friend prepared to climb Mt. Everest. They had to train hard! They watched videos about Mount Everest and learned all they could about the big mountain. They jogged daily to have strong legs. They meditated and did yoga to have strong minds for the challenge. After only three months, their trainers felt she and her friend were ready.

Their trainers sent them to Tibet and the base of Mount Everest alone. Mount Everest sits on the border of Nepal and Tibet. Nepal has a rule that no one younger than 16 can climb the mountain. Nepal had this rule because the climb is so difficult! Tibet did not have this rule. Poorna was only 13. So, Poorna, her friend, and their guides had to climb Mount Everest from the Tibetan side of the mountain, the more difficult way.

Malavath Conquers Everest!

The climb to the top was very hard. While Poorna and her team were climbing, an avalanche killed 16 people. As Poorna climbed, she saw many dead bodies. This was a reminder that every year strong adults die while trying to reach the top of the mountain. But Poorna didn't get scared and she didn't give up. "The training I received helped me **overcome** my fear. I never thought of giving up," she said.

Finally, Poorna reached the top of Mount Everest. She stuck the Indian flag in the ground at the **summit**. She looked around at the view from the highest point in the world. "It was beautiful," she said. "All around me were mountains." Poorna had climbed Mount Everest. This was an incredible thing by itself. When you think that she was a poor girl from a tiny village, and only 13—the act of climbing Everest is superhuman. She described her **motivation** like this: "I wanted to prove that girls could do anything and that students with no money could do anything."

Dangers of Everest

- No Oxygen
- Avalanches
- Extreme wind and cold
- No possiblity of rescue

View of Everest from Tibet taken by Luca Galuzzi.

Map showing the location of Everest.

Figure D. Sample text used during intervention.
This Text was part of the SuperKids unit.

Appendix E

Table E

Sample Based Measure Correlations at Pre-/Posttreatment.

Measure	1	2	3	4	5	6	7	9	10
1. Knowledge	**0.12**	0.45***	0.23*	0.17	0.17	0.13	-0.05		
2. Main Idea	0.03	**0.28**	0.10	-0.07	0.25*	0.11	-0.05		
3. Near RC	-0.09	-0.01	**0.30**	0.48***	0.43***	0.44***	-0.01		
4. Mid RC	0.06	0.00	0.25*	**0.30**	0.40***	0.36**	-0.08		
5. Gates T	-0.01	-0.03	0.31**	0.00	**0.27**	0.63***	0.21		
6. Gates S	-0.10	0.00	0.52***	0.23*	0.45***	**0.61**	0.03		
7. TOWRE SS	-0.13	-0.09	0.01	-0.10	0.03	-0.02	**0.64**		
9. WASI2 Vocab T	0.06	-0.10	0.02	-0.13	0.29**	0.02	0.17		
10. WASI2 Matrix T	-0.01	0.05	-0.01	-0.04	0.02	0.08	0.13	0.00	

Note. Pretreatment correlations below the diagonal; posttreatment correlations are above. Pre/Post correlations for the full sample are on the diagonal. Knowledge is a researcher created measure that assess the vocabulary and concepts taught during intervention. Main Idea is a researcher created measure that tests students' ability to create Main Idea statements for paragraphs. Near RC is a researcher created near transfer test of reading comprehension. Mid Transfer RC is a researcher created measure of Reading Comprehension that includes content not taught in intervention. Gates T is the raw score on the 4 expository passages of the Gates-MacGinite Reading Test-4 - Comprehension, Form t (MacGinite, MacGinite, Maria, & Dreyer, 2000). Gates S is the raw score on the Gates-MacGinite Reading Test-4 Comprehension, full form S. TOWRE SS is the Test of Word Reading Efficiency-2, Sight Word Efficiency subtest standard score (Torgeson, Wagner, & Rashotte, 2012). WASI2 Vocab T and Matrix T are the T scores obtained from the Vocabulary and Matrix Reasoning subtests of the Weschler Abreviated Scales of Intelligence-2 (Weschler, 2011) The WASI2 was administered at pretreatment only.

$*p<.05$; $**p<.01$; $***p<.001$

Appendix F

Table F

	Knowledge	Main Idea	Near RC	Mid RC	Gates T	Gates S
School	0.01	0.00	0.00	0.00	0.00	0.00
Classroom	0.00	0.05	0.00	0.17	0.00	0.00
Pair ID	0.74	0.30	0.00	0.00	0.09	0.00
Individual Error Treatment	0.17	0.45	0.49	0.27	0.43	0.52
Individual Error Control	0.08	0.20	0.51	0.56	0.48	0.48

Note. Knowledge is a researcher created measure that assess the vocabulary and concepts taught during intervention. Main Idea is a researcher created measure that tests students' ability to create Main Idea statements for paragraphs. Near RC is a researcher created near transfer test of reading comprehension. Mid RC is a researcher created measure of Reading Comprehension that includes content not taught in intervention. Gates T is the raw score on the 4 expository passages of the Gates-MacGinite Reading Test-4 - Comprehension, Form t (MacGinite, MacGinite, Maria, & Dreyer, 2000). Gates S is the raw score on the Gates-MacGinite Reading Test-4 Comprehension, full form S.

Appendix G

Table G

Multilevel Model Results for Subsets of Questions on the Near Transfer Reading Comprehension Test

	Main Idea Questions				Theme Questions				Factual Questions				Inference Questions			
Fixed Effect	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*	*Estimate*	*SE*	*t*	*p*
Intercept, γ_{000}	0.19	2.57	0.08	0.94	2.58	2.10	1.23	0.22	1.69	2.83	0.60	0.56	8.01	2.80	2.86	0.01
Pretreatment Score, γ_{100}	0.13	0.14	0.96	0.34	0.17	0.15	1.17	0.25	0.16	0.14	1.12	0.27	0.14	0.15	0.95	0.34
Pretreatment TOWRE SS, γ_{200}	0.02	0.03	0.64	0.53	-0.01	0.02	-0.47	0.64	0.01	0.03	0.32	0.75	-0.05	0.03	-1.94	0.06
TRT vs Control, γ_{300}	0.00	0.31	0.01	1.00	0.28	0.27	1.07	0.29	0.07	0.36	0.18	0.86	0.90	0.34	2.63	0.01
Random Effect																
Teacher (τ_u)	0.00				0.14				0.00				0.00			
Pair (τ_r)	0.02				0.23				0.05				0.05			
Residual (treatment) (σ_{12})	1.78				0.79				2.84				2.35			
Residual (control) (σ_{22})	1.66				1.14				1.82				1.95			

Note. The Near Transfer Reading Comprehension test is a researcher created near transfer test of reading comprehension. There were 6 main idea questions, 4 theme questions, 6 factual questions, and 8 inference questions on the Near RC measure. Table represents results for cross-classified hierarchical linear models with random effects for teacher and pair.

REFERENCES

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

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2010). Reading comprehension instruction for students with learning disabilities, 1995—2006: A meta-analysis. *Remedial and Special Education*, *31*(6), 423-436. https://doi.org/10.1177/0741932509355988

Catts, H. W., & Kamhi, A. G. (2017). Prologue: Reading comprehension is not a single ability. *Language, Speech, and Hearing Services in Schools*, *48*(2), 73-76. https://doi.org/10.1044/2017_LSHSS-16-0049

Cervetti, G. N., Jaynes, C. A., & Hiebert, E. H. (2009). Increasing opportunities to acquire knowledge through reading. In E. H. Hiebert (Ed.) *Reading more, reading better* (79-100). New York City: Guilford Press.

Clemens, N. & Fuchs, D. (2020). *Commercially-developed tests of reading comprehension: Gold standard or fool's gold.* Unpublished manuscript.

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of "quick fix" interventions for children with reading disability? *Scientific Studies of Reading*, *18*(1), 55-73. https://doi.org/10.1080/10888438.2013.836200

Connor, C. M., Phillips, B. M., Kim, Y. G., Lonigan, C. J., Kaschak, M. P., Crowe, E., Dombek, J., & Al Otaiba, S. (2018). Examining the efficacy of targeted component interventions on language and literacy for third and fourth graders who are at risk of comprehension difficulties. *Scientific Studies of Reading, 22*(6), 462-484. https://doi.org/10.1080/10888438.2018.1481409

Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology*, *80*(2), 131. https://doi.org/10.1037/0022-0663.80.2.131

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515-528. https://doi.org/10.1037/a0027182

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research*, *79*(1), 262-300. https://doi.org/10.3102/0034654308325998

Elleman, A. M. (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*, *109*(6), 761. https://doi.org/10.1037/edu0000180

Elleman, A. M., & Compton, D. L. (2017). Beyond comprehension strategy instruction: What's next? *Language, Speech, and Hearing Services in Schools*, *48*(2), 84-91. https://doi.org/10.1044/2017_LSHSS-16-0036

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading, 10*(3), 323-330. https://doi.org/10.1207/s1532799xssr1003_7

Fuchs, D., & Fuchs, L. S. (2015). Rethinking service delivery for students with significant learning problems: Developing and implementing intensive instruction. *Remedial and Special Education*, *36*(2), 105-111. https://doi.org/10.1177/0741932514558337

Fuchs, D., Fuchs, L. S., Thompson, A., Otaiba, S. A., Yen, L., Yang, N. J., Braun, M., & O'Connor, R. E. (2001). Is reading important in reading-readiness programs? A randomized field trial with teachers as program implementers. *Journal of Educational Psychology*, *93*(2), 251. https://doi.org/10.1037/0022-0663.93.2.251

Fuchs, D., Hendricks, E., Walsh, M. E., Fuchs, L. S., Gilbert, J. K., Zhang Tracy, W., Patton, S., Kim, W., Elleman, A., & Peng, P. (2018). Evaluating a multidimensional reading comprehension program and reconsidering the lowly reputation of tests of near transfer. *Learning Disabilities Research & Practice, 33*(1), 11-23. http://dx.doi.org/10.1111/ldrp.12162

Gajria, M., Jitendra, A. K., Sood, S., & Sacks, G. (2007). Improving comprehension of expository text in students with LD: A research synthesis. *Journal of Learning Disabilities*, *40*(3), 210-225. https://doi.org/10.1177/00222194070400030301

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*(2), 279-320. https://doi.org/10.3102/00346543071002279

Gilbert, J. K., Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Barquero, L. A., & Cho, E. (2013). Efficacy of a first-grade responsiveness-to-intervention prevention model for struggling readers. *Reading Research Quarterly*, *48*(2), 135-154. https://doi.org/10.1002/rrq.45

Gilmour, A. F., Fuchs, D., & Wehby, J. H. (2019). Are students with disabilities accessing the curriculum? A meta-analysis of the reading achievement gap between students with and without disabilities. *Exceptional Children*, *85*(3), 329-346. https://doi.org/10.1177/0014402918795830

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202. https://doi.org/10.3758/BF03195564

Hendricks, E. & Fuchs, D. (in press). Are individual differences in response to intervention influenced by the methods and measures used to define response? Implications for identifying children with learning disabilities. *Journal of Learning Disabilities*.

Hernandez, D. (2011) *Double Jeopardy: How third-grade teading skills and poverty influence high school graduation*. Report prepared for the Annie E. Casey Foundation.

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, *47*(2), 125-135. https://doi.org/10.1177/0022219412439326

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. https://doi.org/10.21236/ADA006655

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*(5), 242-252. https://doi.org/10.3102/0013189X14539189

Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes. *Journal of Educational Psychology*, *109*(7), 889. https://doi.org/10.1037/edu0000181

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

Mastropieri, M. A., Scruggs, T. E., Bakken, J. P. & Whedon, C. (1996). Reading comprehension: A synthesis of research in learning disabilities. In T. E. Scruggs & M. A. Mastropieri (Eds.) *Advances in learning and behavioral disabilities: Intervention research* (vol. 10, Part B, pp. 201-227). Greenwich, CT: JAI Press.

McMaster, K. L., Espin, C. A., & Van Den Broek, P. (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice*, *29*(1), 17-24. https://doi.org/10.1111/ldrp.12026

Meneghetti, C., Carretti, B., & De Beni, R. (2006). Components of reading comprehension and scholastic achievement. *Learning and Individual Differences*, *16*(4), 291-301. https://doi.org/10.1016/j.lindif.2006.11.001

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups*. Report prepared for National Institute for Child Health and Human Development. Retrieved from https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

O'Connor, R. E., & Fuchs, L. S. (2013). Responsiveness to intervention in the elementary grades: Implications for early childhood education. In *Handbook of response to intervention in early childhood*, by Buysse, Virginia (ed.); Peisner-Feinberg, Ellen S (ed.). 41-55, Chapter xvi, 468 Pages. Baltimore, MD: Paul H Brookes Publishing Co.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, *8*(3), 293-316. https://doi.org/10.1016/0361-476X(83)90018-8

Pickering, S., & Gathercole, S. (2001). *Working Memory Test Battery for Children.* London: The Psychological Corporation.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *64*(4), 479-530. https://doi.org/10.3102/00346543064004479

Scammacca, N. K., Roberts, G. J., Cho, E., Williams, K. J., Roberts, G., Vaughn, S. R., & Carroll, M. (2016). A century of progress: Reading interventions for students in grades 4–12, 1914–2014. *Review of Educational Research*, *86*(3), 756-800. https://doi.org/10.3102/0034654316652942

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*(4), 369-390. https://doi.org/10.1177/0022219413504995

Schwanenflugel, P. J., Meisinger, E. B., Wisenbaker, J. M., Kuhn, M. R., Strauss, G. P., & Morris, R. D. (2006). Becoming a fluent and automatic reader in the early elementary school years. *Reading Research Quarterly*, *41*(4), 496-522. https://doi.org/10.1598/RRQ.41.4.4

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). *Improving reading comprehension in kindergarten through 3rd grade: IES practice guide.* NCEE 2010-4038. *What Works Clearinghouse*.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

StataCorp. (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.

Stenner, A. J., Smith M. A., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305-316. https://doi.org/10.1111/j.1745-3984.1983.tb00209.x

Sterba, S. K. (2017). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research, 27*(4), 425-436. http://dx.doi.org/10.1080/10503307.2015.1114688

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., Clevenger, W., Wasdell, M., McCleary, R., Lakes, K., & Wigal, T. (2012). Categorical and dimensional definitions and evaluations of symptoms of ADHD: History of the SNAP and the SWAN rating scales. *The International Journal of Educational and Psychological Assessment*, 10 (1), 51.

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

Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., Stancavage, F., Durno, D., Javorsky, R. & Haan, C. (2006). *National assessment of Title I: Interim report. Volume II: Closing the reading Gap: First year findings from a randomized trial of four reading interventions for striving readers.* Washington, DC: National Center for Education Evaluation and Regional Assistance.

Walsh, M. E. (2020) *Meta-analysis of reading comprehension interventions for struggling third-grade readers*. Doctoral Qualifying Exam, presented to dissertation committee at Vanderbilt University.

Walsh, M. E., Fuchs, D., Patton, S. A., & Fuchs, L. S. (2019). *Developing a comprehensive comprehension intervention for struggling third grade readers: A pilot study*. Poster presentation. Pacific Coast Research Conference, San Diego, CA.

Walsh, M. E., Fuchs, D., Patton, S.A., Gilbert, J., (2017) *Differential efficacy of reading comprehension intervention for third and fourth grade students.* Poster presentation. Pacific Coast Research Conference, San Diego, CA.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review*, *36*(4), 541-561.

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of tier 2 type reading interventions in grades K-3. *Educational Psychology Review*, *28*(3), 551-576. https://doi.org/10.1007/s10648-015-9321-7

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. https://doi.org/10.1007/s11145-009-9179-5

Webb, N. M. (1989). Peer interaction and learning in small groups. *International journal of Educational Research*, *13*(1), 21-39. https://doi.org/10.1016/0883-0355(89)90014-1

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

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence* (2nd ed.). San Antonio, TX: Pearson. https://doi.org/10.1037/t15171-000

What Works Clearinghouse. (2008). *What Works Clearinghouse Procedures Handbook, Version 4.0*. Institute of Educational Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf