Title

The Effects of One versus Two Years of Intensive Reading Intervention Implemented with Late Elementary Struggling Readers

Permalink
https://escholarship.org/uc/item/17q9x6ms

Journal
Learning Disabilities Research and Practice, 33(1)

ISSN
0938-8982

Authors
Miciak, Jeremy
Roberts, Garrett
Taylor, W Pat
et al.

Publication Date
2018-02-01

DOI
10.1111/ldrp.12159

Peer reviewed

Learning Disabilities Research & Practice, *00*(0), 1–13

DOI: 10.1111/ldrp.12159

LEARNING DISABILITIES
RESEARCH & PRACTICE

The Effects of One versus Two Years of Intensive Reading Intervention Implemented with Late Elementary Struggling Readers

Jeremy Miciak
The University of Houston

Garrett Roberts
The University of Denver

W. Pat Taylor
The University of Houston

Michael Solis
University of California Riverside

Yusra Ahmed
The University of Houston

Sharon Vaughn
The University of Texas at Austin

Jack M. Fletcher
The University of Houston

We examined the effectiveness of a researcher-provided reading intervention with 484 fourth graders with significant reading difficulties. Students were randomly assigned to one year of intervention, two years of intervention, or a business-as-usual comparison condition (BAU). Students assigned to two years of intervention demonstrated significantly greater gains in reading fluency compared to students who received one year of intervention and the BAU group. Students in both the one- and two-year groups demonstrated similar and significantly larger gains in word reading in comparison to the BAU group. There were no statistically significant differences between the three groups on standardized measures of reading comprehension. We discuss these results in the context of research with late elementary and secondary students targeting reading comprehension.

Late elementary is a critical period in reading development. First, it represents an educational benchmark by which time students are expected to read extensively in texts that contain challenging vocabulary and increasingly complex content, not just during reading instruction, but across subject areas. Second, the content and focus of reading instruction shift. In late elementary, reading instruction focuses primarily on comprehension skills and strategies including analysis across texts, understanding genres, and building academic vocabulary, and shifts away from explicit instruction in foundational reading skills. As a result, students who enter late elementary grades lacking foundational reading skills may face compounding academic challenges in subsequent schooling (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Lesnick, Goerge, Smithgall, & Gwynne, 2010).

READING INTERVENTIONS IN LATE ELEMENTARY GRADES

When delivered with sufficient intensity and dosage to remediate foundational skills, interventions in early elementary grades have demonstrated the potential to alter future educational trajectories, particularly with regard to word-level reading skills (Blachman et al., 2014; Wolff, 2016). However, interventions targeting students in late elementary grades have not demonstrated similarly robust findings, particularly for reading comprehension. Intervention studies conducted with struggling readers in these age ranges have typically yielded modest to minimal effects, and few

Requests for reprints should be sent to Jeremy Miciak University of Houston. Electronic inquiries should be sent to Jeremy.Miciak@times.uh.edu.

students progress sufficiently within these interventions to approach reading levels commensurate with their typically reading peers (Flynn, Zheng, & Swanson, 2012; James-Burdumy et al., 2012; Scammacca, Roberts, Vaughn, & Stuebing, 2015bb).

Despite the need for robust interventions in these grades, comparatively few rigorous studies have evaluated the effects of intensive reading interventions with late elementary students. In one notable study, a large-scale randomized control trial conducted with over 10,000 5th grade students in Title I schools failed to demonstrate statistically significant, positive effects for four supplemental reading comprehension interventions implemented with struggling readers (James-Burdumy et al., 2012). Each of the four intervention programs focused on improving comprehension of informational text using evidence-based practices, and followed an explicit instructional routine of teacher modeling and scaffolding with guided practice to reach independent application of strategies. Post-intervention comparisons of the four intervention groups and a comparison group yielded no statistically significant positive effects on reading comprehension outcomes for any group after a full year of implementation, although one intervention yielded a small effect size of 0.22 when teachers with prior experience with the curriculum provided the intervention.

Such results are consistent with previous syntheses and meta-analyses, which report that interventions beyond early elementary often yield small to moderate positive effects (if any) on standardized measures of reading comprehension when rigorous designs are employed (Scammacca et al., 2015ab; Solis et al., 2012; Wanzek, Wexler, Vaughn, & Ciullo, 2010). For example, Scammacca and colleagues conducted a meta-analysis of reading intervention studies focused on grades 4–12. Across 82 study-wise effects, the meta-analysis yielded a mean meta-analytic effect (*d*) of 0.49 for all measures; however, the mean effect size was 0.21 for standardized measures of reading. Additional analyses found no significant moderating effect for grade-level (grades 4–5, 6–8, and 9–12). A separate, selective meta-analysis of studies evaluated the effects for experimental studies of reading interventions implemented with struggling readers in grades 5–9 on norm-referenced, standardized measures of reading only (Flynn et al., 2012). Results were generally consistent with those of Scammacca et al.; only 10 studies met inclusion criteria, and these studies generally yielded small to moderate effect sizes on standardized measures.

STUDY PURPOSE

The present study was designed to address several critical issues highlighted in previous meta-analyses and syntheses. First, the study was conducted as a randomized control trial to address the paucity of large-scale, experimental studies implemented with late elementary students with reading difficulties. Second, the study utilized standardized measures of reading, including decoding, fluency, and comprehension. Third, the study was conducted over two years, with students initially randomized to one of three conditions: (1) two years of researcher-provided treatment; (2) one year of researcher-provided treatment; and (3) a business as usual comparison condition, which included school-provided interventions. The multi-year intervention protocol allowed us to evaluate an intervention of greater duration and dosage than is typically reported in research at these grades. Fourth, the study included a comprehensive intervention protocol that addressed foundational reading skills and more complex language and text-based processes to facilitate differentiated instruction according to student needs. Rather than focusing on cognitive strategy instruction, we conceptualized the treatment to support building knowledge structures from text-based representations followed by integration of the ideas through discussion to support better understanding (e.g., Beck & McKeown, 2006; Gersten, Baker, Smith-Johnson, Dimino, & Peterson, 2006; McKeown, Beck, & Blake, 2009). Thus, the study addresses an important gap in research on reading comprehension by moving away from short-term, strategy-specific studies that incorporate proximal measures that do not or only partially tap the latent construct of reading comprehension, as theoretically understood (Compton, Miller, Elleman, & Steacy, 2014; Hirsch, 2006).

Intervention development was primarily guided by the Simple View of Reading (SVR; Gough & Tunmer, 1986; Hoover & Gough, 1990). Together, the following three components were organized to support all domains of the SVR: (1) *word reading* (automaticity of sight and decodable words), (2) *world knowledge* (key concepts, vocabulary, and language supports), and (3) *text-processing strategies* (including mental models and inference-making). Components 2 and 3 targeted the language domain, and component 1 targeted the word reading component within the SVR framework. Component two was also derived from models of text-processing that focus on the integration of ideas from text to promote better understanding (e.g., Kintsch, 1974; van den Broek, Young, Tzeng, & Linderholm, 1998).

Using a randomized control trial design, we compared fourth- and fifth-grade reading outcomes for students with severe reading difficulties in three conditions: students who received two years of researcher-provided, intensive intervention (fourth and fifth grades); students who received only one year of researcher-provided, intensive intervention (fourth grade only); and students who received a "business-as-usual" comparison condition. All participants were followed for two years.

We hypothesized that:

1. Students assigned to the researcher-provided intervention in fourth grade (both the 1-year and 2-year intervention groups) will outperform students in the BAU comparison condition at the end of fifth grade on all reading outcomes.
2. Students assigned to the researcher-provided intervention for two years (4th and 5th grades) will outperform students who received one year of intervention and students in the BAU comparison condition at the end of 5th grade on all reading outcomes.

3. Students assigned to one year of researcher-provided intervention (4th grade) will continue to outperform students in the BAU comparison condition at the end of 5th grade, although these differences will diminish over time.

METHOD

Participants

School Sites

Participants for this study were drawn from 17 participating schools, distributed across two sites. One site included eight schools in a large, urban district in the southwestern United States. The other site included nine schools distributed across two districts: a suburban district in a large urban area in the southwestern United States, and an ex-urban district located outside the same urban area. All schools were rated academically acceptable by the state education agency, and included a large number of students receiving free and/or reduced price lunch.

Selection of Participants

We selected participants based on a whole-school screening process during students' fourth-grade year. All students at participating schools enrolled in general education English language arts classes completed the Gates–MacGinitie Reading Test (MacGinitie et al., 2000) in the fall of that year. Students who received a standard score equal to or less than 85 were eligible for participation.

Student Participants

A priori estimates of statistical power indicated that a sample size of 420 students, distributed across three assignment groups, was sufficient to detect small to moderate effects sizes at power <.80. School accountability data were used to estimate the screening yield at each campus, and to guide the number of screened schools. Screening yielded 484 eligible students, who were all included as a protection against attrition, and to yield greater statistical power. Eligible students ($N = 484$) were 45 percent female, 49.7 percent LEP, 23.6 percent SPED, 68 percent Hispanic, 23 percent African American, 8 percent White, and 1 percent Other or Unknown. As a proxy for evaluating economic disadvantage, 13 percent did not receive free or reduced lunch, 5 percent were receiving free or reduced lunch on a temporary basis, and 82 percent were receiving free or reduced lunch. The average score on the Gates-MacGinitie test for eligible students was 77.2 (SD = 6.1).

Students were blocked by school and English learner (EL) status (EL or not-EL based on extant school records) and assigned in a 1:1:1 ratio to three conditions: (1) a two-year treatment protocol (two-year) that spanned from fall of year 1 (fourth grade) to spring of year 2 (fifth grade, except in cases of retention); (2) a one-year treatment protocol (one-year) that spanned from fall of year 1 to spring of year 1 (fourth grade) and included an untreated (but tested) BAU comparison condition in year 2; and (3) a BAU comparison condition in both years 1 and 2. After randomization there were 162 students in the two-year group, 161 in the one-year group, and 161 in the BAU group. There was no significant relation between treatment group and gender ($\chi^2 = 2.10$, $df = 2, p = .35$), treatment group and race ($\chi^2 = 5.48, df = 6$, $p = .48$), or treatment group and economic disadvantage ($\chi^2 = 6.61$, $df = 8$, $p = .58$). Additionally, an omnibus MANOVA indicated that there were no significant differences between treatment groups on pretest measures, Wilks' Lambda = 0.96, $F(10,864) = 1.83, p > .05$.

Attrition

Across the two years, 73 students were unavailable for posttest. Twenty students withdrew from the study, eight were removed from the study by school request, 37 moved from the participating school, one refused to participate, one missed testing due to extended absence, and six students were lost due to clerical errors. There was no significant difference on the Gates-MacGinitie test between those who attrited and those who did not attrit ($p = .09$). There was no significant relation between those who attrited and site ($\chi^2 = 1.36, df = 1, p = .24$).

Measures

Decoding and Spelling

We assessed word reading accuracy with the Letter-Word Identification subtest of the Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001). Published test-retest reliabilities for the Letter-Word Identification subtest for students aged 8–13 range from .89 to .96 (median $r = .93$; McGrew, Schrank, & Woodcock, 2007). We assessed spelling with the Spelling subtest of the WJ-III (Woodcock et al., 2001). Test-retest reliabilities for students aged 8–13 range from .87 - .89 (median $r = .89$; McGrew et al., 2007). We report standard scores for descriptive purposes; however, vertically aligned Rasch (W) scores are utilized in all analyses.

Fluency

We assessed single-word fluency with subtests from the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999). The Sight Word Efficiency subtest assesses real world reading, whereas Pseudoword Reading Efficiency assesses fluent reading with phonetically regular nonsense words. Alternate form reliability coefficients for both subtests are high, in excess of .90 (Torgesen et al., 1999). We utilized standard scores for all analyses on the TOWRE, because vertically aligned scores are not available.

Comprehension

We assessed reading comprehension with two tests. The Gates–MacGinitie (MacGinitie et al., 2000) is an untimed, group-administered assessment of reading comprehension. The task requires students to read expository and narrative passages ranging in length from 3 to 15 sentences and answer 3–6 multiple choice questions related to the passage. Published internal reliability coefficients with students in grades 4 and 5 range from .91 to .92. We utilized the Form S for grade 4 at both time points in year 1, and Form S for grade 5 at both time points in year 2. We report standard scores for descriptive purposes; vertically aligned extended scale scores were utilized for all analyses. The WJ-III Passage Comprehension subtest (Woodcock et al., 2001) is a cloze assessment of reading in which the student reads a short passage of text and provides the missing words. Published test-retest reliabilities for children aged 8–13 range from .80 to .92 (median $r = .86$; McGrew et al., 2007). We report standard scores for descriptive purposes; however, vertically aligned W scores are utilized in all analyses.

Intervention Procedures

Tutors and Training

The intervention was taught by 19 tutors (18 female) in year one and 12 (11 female) in year two. The research team hired, trained, and supervised all tutors. All tutors held at least a Bachelor's degree, and all were experienced working with children and in schools. Twelve of 27 tutors (44 percent) held a relevant teaching credential (e.g., K-8 or special education), and 11 (41 percent) held a Master's degree in education.

Tutors were trained prior to intervention implementation by members of the research team. Experienced members of the research team trained tutors prior to intervention implementation. Training consisted of one day (8 hours) of professional development that introduced the components of the intervention and two half-days for tutor practice with guided feedback. All tutors were provided classroom coaching sessions for the first couple of weeks, in addition to ongoing supports as needed throughout the duration of intervention. Progress monitoring data were also used to determine tutors/groups that needed additional supports.

During year one of the study (fourth grade), the intervention instruction included vocabulary, word study, and text reading (Vaughn, Solis, Miciak, Taylor, & Fletcher, 2016). The same instructional components were included during year two (fifth grade) with the addition of a self-regulation component. Participants received small group instruction (4 to 6 per group) for 30- to 40-min sessions five times per week for approximately 16 weeks. Lessons were structured around two-week units aligned to social studies (4th grade) and science (5th grade) standards within this state. Each lesson contained word study and passage reading components. Vocabulary instruction and self-regulation instruction were provided for six and three days, respectively. The final day of the unit allotted time to re-read text, and for curriculum-based measurements to monitor progress.

Text-based Reading

The text-based reading component incorporated two levels of text: stretch text (three days) and fluency text (six days). Stretch texts were grade-level science and social studies texts that had been adapted for readability. Stretch texts were read utilizing choral reading, partner reading, or independent reading. The texts included stopping points every three to four paragraphs. At each stopping point, students summarized the meaning of the text in their own words, and tutors asked questions about the text. When students had difficulty answering questions, tutors scaffolded the activity by identifying the specific section(s) of text necessary to accurately answer the question. Tutors were encouraged to differentiate instruction for students at different reading levels by utilizing different reading formats, varying the length of text read for repeated reading, and providing additional scaffolds.

Fluency texts were science- and social studies-themed texts from QuickReads (Hiebert, 2003), which were designed to facilitate comprehension through a reduction in the number of unfamiliar words. Tutors selected fluency texts that were appropriate for students' reading level. Prior to reading, students previewed text and identified unknown words. The instructional routine began with teacher-modeled or choral reading followed by independent or partner reading. After reading instruction, students completed a "Does it Make Sense?" activity that required students to analyze sentences to identify whether the syntax and semantics were internally congruent. For sentences that did not make sense, students identified the word or phrase that introduced the incongruity.

Word Study

Daily word study addressed phonics skills and sight word reading with a goal of automaticity. Students practiced reading high-frequency sight words and multi-syllable words at the word, phrase, and sentence level. Tutors identified appropriate word lists for individual students based on their word reading skill. The instructional routine included tutor modeling of automatic word reading, as well as reading and re-reading the word lists with peer and tutor feedback. When students mastered a list, they progressed to increasingly difficult word lists.

Vocabulary

Explicit vocabulary instruction was delivered for approximately five mins on days that featured fluency-text instruction prior to text reading. For each two-week unit, five vocabulary words directly related to the text-based readings were identified. Instruction included simplified definitions, word use in context, and discussion questions designed to engage students in understanding the meaning of the words.

Self-regulation

The self-regulation instructional component was included as a mechanism to focus student attention towards improving vocabulary knowledge. Lessons utilized a self-monitoring sheet for goal setting on the number of vocabulary words to be learned, as well as to identify and reflect on their ability to stay motivated and engaged. Students identified the number of new vocabulary words they would learn as well as identifying whether could meet the predetermined attribution statements of believing in oneself, identifying ways to overcome difficulties, and perseverance. Following each lesson, students' goals were evaluated through the assessment of the number of vocabulary words learned and a self-reflection on their ability to meet the attribution statements' expectations.

Business as Usual Comparison Condition

Students in the BAU comparison condition also received reading interventions provided by school personnel. To establish a better understanding of the interventions being provided, we interviewed school personnel and asked them to complete the Additional Reading Interventions Inventory (described below) to determine the amount and type of ongoing interventions being provided. Multiple interventions were described by school personnel, and we categorized these interventions broadly as test preparation, basic word reading intervention, fluency interventions, inclusion support, and RTI/resource support. However, it was not feasible to conduct direct observations of these interventions, and it is uncertain whether instruction matched what was described, and to what extent school-provided intervention activities overlapped with the researcher-provided intervention.

Fidelity of Implementation

We treat fidelity of implementation as a multi-dimensional construct, measuring the extent to which the programs were implemented as planned. Dane and Schneider (1998) identified five components of treatment fidelity that have been previously used in educational research, including (1) adherence, (2) quality, (3) dosage, (4) differentiation with the comparison condition, and (5) treatment receipt. We report data related to the first four components, but do not report data on treatment receipt. In previous educational literature, treatment receipt has been operationalized with behaviors such as student engagement or practice opportunities, measures that we were not able to collect due to resource limitations.

Instruments and Procedures

Adherence to the intervention protocols and quality of implementation were evaluated for each tutor across eight lessons, using the Fidelity of Implementation Instrument. All lessons were audio-recorded. Recordings were blocked by reading group, and eight were randomly selected for each tutor. Prior to individual coding, all fidelity coders independently coded two randomly selected lessons to evaluate inter-rater reliability with a "gold standard" score (Gwet, 2001). The process was repeated until the coder achieved agreement >90 percent, at which point the coder was allowed to complete independent coding. To protect against rating drift, a second reliability check was conducted following the same procedures after the coder had completed 50 percent of the lessons assigned to her.

The Additional Reading Interventions Inventory (ARI), a tool developed by the research team, was utilized to evaluate the amount and nature of school-based reading interventions for students in the researcher-provided reading intervention and the comparison condition. The ARI is a survey instrument in which classroom teachers report all additional reading services that each student has received. It was administered in spring of year 1, winter of year 2, and spring of year 2. The ARI includes items documenting the frequency and duration of school-based interventions, as well as the nature of those interventions.

Adherence

Adherence evaluates to what extent the components of the intervention were implemented as designed. Adherence was evaluated via 4-point Likert-type items for each component of the intervention, including the specific steps to deliver the instructional component (1 = not present, 2 = inconsistently present, 3 = mostly present, 4 = always present). Across all tutors and components, adherence was high ($M = 3.6$, $SD = 0.76$). Data reveal very little variability among component means (range = 3.2–4.0), year means (range = 3.6–3.6), and tutor means (range = 3.0–4.0).

Quality

Quality of implementation evaluates how well the intervention, viewed holistically, was implemented. Quality of implementation was evaluated via a 5-point Likert-type scale for each coded lesson, with a score of 5 indicating highest quality, 3 indicating average quality, and 1 lowest quality. Across all tutors, quality was generally average to high ($M = 3.6$, SD = 0.60).

Dosage

We define dosage as the total time in which the individual student participated in the researcher-provided intervention. Across students that did not attrit, the mean hours in intervention during the first year were 42.41 (SD = 6.71) for the one-year group and 41.94 (SD = 7.46) for the two-year group. Students in the two-year group received 29.43 hours (SD = 8.27) of intervention during the second year. Dosage did not vary by site for year 1 (Site 1: $M = 41.35$, SD = 7.23; Site 2: $M = 42.95$, SD = 6.89), whereas there were site differences for year 2 (Site 1: $M = 27.60$, SD = 7.88; Site 2: $M = 31.36$, SD = 8.32; $t = -2.33$, $df = 99$, $p = .02$) and a significant difference by year (Year 1: $M = 41.86$, SD = 8.08; Year 2: $M = 29.43$, SD = 8.32; $t = 12.71$, $df = 326$, $p < .001$). These differences by site and year are due to a school change in one participating district in which students in fifth grade transitioned to middle school with subsequent scheduling limitations at those middle schools.

Differentiation with Comparison Condition

The ARI is a summary of teacher-reported reading instruction occurring outside the normal language arts block. Across students and years, the mean amount of ARI was 49.11 hours (SD = 69.37) for the one-year treatment group, 33.45 hours (SD = 57.75) for the two-year group, and 50.07 hours (SD = 66.82) for the BAU group. The total additional reading instruction (T-ARI) is the sum of teacher-provided and researcher-provided instruction. Across students and years, the mean amount of T-ARI was 85.14 hours (SD = 75.90) for the one-year group, 88.48 hours (SD = 68.06) for the two-year group, and 50.07 hours (SD = 66.82) for the BAU group.

Data Analysis

A repeated-measures multivariate analysis of variance (MANOVA) was conducted to evaluate the hypotheses. Treatment condition served as the between-subjects factor, and time served as the within-subjects factor. We included the two blocking variables—EL status and school—as covariates to account for the effects of clustering. We evaluated results against $\alpha = .05$ in the initial MANOVA. Control of α due to multiple testing in follow-up analyses was achieved by implementing the false discovery rate (FDR) method of control (Benjamini & Hochberg, 1995). Follow-up analyses comprised univariate repeated-measures analyses for each outcome measure and single degree of freedom contrasts to evaluate group differences for linear and quadratic slopes.

RESULTS

Average scores for each outcome are reported by time point for each group in Table 1. Table 2 shows Cohen's *d* for group comparisons by test and time point. In all cases, the comparison is the group with more intervention minus the group with less intervention.

The repeated measures MANOVA tested for effects of outcome, outcome by group, outcome by school, outcome by ELL status, outcome by time, outcome by group by time, outcome by school by time, and outcome by ELL status by time. The effect of outcome was significant (Wilks' $\lambda < .001$, $F(5, 286) = 104363, p < .0001$). The effect of outcome by group was also significant (Wilks' $\lambda = .93$, $F(10, 572) = 2.12$, $p = .02$), as was the effect of outcome by school (Wilks' $\lambda = .68$, $F(80, 1382) = 1.76, p = .0086$. The effect of outcome by ELL status was significant (Wilks' $\lambda = .75$, $F(5, 286) = 18.74$, $p < .0001$). The effect of outcome by time was significant (Wilks' $\lambda = .20$, $F(10, 281) = 112.48$, $p < .0001$), as was the effect of outcome by time by group (Wilks' $\lambda = .86, F(20, 562) = 2.10, p = .0036$). The effect of outcome by school by time was also significant (Wilks' $\lambda = .42, F(160, 2419) = 1.63, p < .0001$). The effect of outcome by ELL status by time was not significant (Wilks' $\lambda = .94$, $F(10, 281) = 1.69, p = .08$).

TABLE 1
Descriptive Statistics by Time Point and Assignment Group

	BAU		*One-year*		*Two-year*	
	N	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*
WJ-III LWID						
Pretest Year 1	149	90.4 (11.01)	148	88.07 (11.91)	148	90.24 (10.47)
Posttest Year 1	136	92.46 (10.78)	131	91.86 (11.25)	137	92.79 (9.54)
Posttest Year 2	105	91.96 (11.76)	117	92.49 (12.27)	114	94.12 (11.54)
TOWRE Sight Word						
Pretest Year 1	150	80.95 (12.07)	148	78.18 (12.41)	147	81.61 (11.75)
Posttest Year 1	136	84.46 (12.09)	131	83.84 (12.54)	137	86.36 (12.11)
Posttest Year 2	105	87.5 (12.81)	117	86.12 (12.45)	114	90.94 (13.25)
WJ-III Spelling						
Pretest Year 1	148	88.11 (10.56)	145	85.55 (10.67)	147	87.23 (9.82)
Posttest Year 1	136	88.85 (12.69)	131	85.95 (14.56)	137	87.6 (12.98)
Posttest Year 2	105	89.12 (14.19)	117	88.78 (13.85)	114	90.48 (12.21)
WJ-III PC						
Pretest Year 1	149	82.77 (8.77)	148	80.19 (9.23)	148	82.55 (8.4)
Posttest Year 1	136	84.81 (8.67)	131	82.34 (9.45)	137	84.36 (8.64)
Posttest Year 2	105	84.5 (9.96)	117	82.43 (9.98)	114	84.9 (8.84)
Gates-MacGinitie						
Pretest Year 1	161	77.14 (6.17)	161	77.01 (6.33)	162	77.49 (5.85)
Posttest Year 1	137	84.53 (8.92)	133	83.18 (7.89)	137	84.93 (8.04)
Posttest Year 2	102	86.53 (9.92)	113	86.24 (8.5)	112	86.1 (8.61)

WJ-III LWID = Woodcock Johnson-Third Edition: Letter Word Identification. TOWRE = Test of Word Reading Efficiency; WJ-III PC = Woodcock Johnson-Third Edition: Passage Comprehension.

TABLE 2
Cohen's *d* for Group Comparisons by Test and Time Point

		Pretest Year 1		*Posttest Year 1*		*Posttest Year 2*	
Test	*Comparison*	*d*	*CI*	*d*	*CI*	*d*	*CI*
WJ-III LWID							
	Two-year - One-year	0.21	(−0.02, 0.44)	0.09	(−0.15, 0.33)	0.14	(−0.11, 0.40)
	Two-year - BAU	−0.08	(−0.31, 0.15)	0.02	(−0.22, 0.26)	0.14	(−0.12, 0.41)
	One-year -BAU	−0.27	(−0.50, −0.05)	−0.07	(−0.31, 0.17)	−0.01	(−0.27, 0.26)
TOWRE Sight Word							
	Two-year - One-year	0.3	(0.07, 0.53)	0.18	(−0.06, 0.42)	0.36	(0.10, 0.62)
	Two-year - BAU	0.03	(−0.19, 0.26)	0.12	(−0.12, 0.36)	0.22	(−0.04, 0.49)
	One-year - BAU	−0.26	(−0.49, −0.04)	−0.07	(−0.31, 0.17)	−0.14	(−0.40, 0.13)
WJ-III Spelling							
	Two-year - One-year	0.16	(−0.07, 0.39)	0.1	(−0.14, 0.34)	0.13	(−0.13, 0.39)
	Two-year - BAU	−0.16	(−0.39, 0.07)	−0.13	(−0.36, 0.11)	0.08	(−0.18, 0.35)
	One-year - BAU	−0.32	(−0.55, −0.09)	−0.23	(−0.47, 0.01)	−0.04	(−0.31, 0.22)
WJ-III PC							
	Two-year - One-year	0.28	(0.05, 0.51)	0.21	(−0.03, 0.45)	0.26	(0.00, 0.52)
	Two-year - BAU	−0.13	(−0.35, 0.10)	−0.11	(−0.35, 0.13)	−0.02	(−0.29, 0.24)
	One-year - BAU	−0.4	(−0.63, −0.17)	−0.32	(−0.56, −0.08)	−0.27	(−0.53, 0.00)
Gates-MacGinitie							
	Two-year - One-year	0.09	(−0.12, 0.31)	0.23	(−0.01, 0.47)	−0.02	(−0.29, 0.24)
	Two-year - BAU	0.07	(−0.15, 0.29)	0.04	(−0.19, 0.28)	−0.05	(−0.32, 0.22)
	One-year - BAU	−0.03	(−0.24, 0.19)	−0.17	(−0.41, 0.07)	−0.03	(−0.30, 0.24)

Note: All comparisons include the greater dosage condition as the minuend; thus, positive effects are consistent with stated hypotheses. WJ-III LWID = Woodcock Johnson-Third Edition: Letter Word Identification. TOWRE = Test of Word Reading Efficiency; WJ-III PC = Woodcock Johnson-Third Edition: Passage Comprehension.

The significant effect of outcome was not of interest because it was likely due to the different metrics of the outcomes. Additionally, this effect did not include group, so it was not useful in evaluating our hypotheses. Similarly, we did not have hypotheses requiring the examination of the effects of outcome by time, outcome by school, outcome by school by time, outcome by ELL status, or outcome by ELL status by time. As a result, the only effect that was subjected to follow-up analyses was the effect of outcome by time by group. This test served to determine which of the outcome measures demonstrated a significant time by group interaction. Separate repeated measures ANOVAs were fit for each outcome measure (Table 3). After running these analyses, we used the MULTTEST procedure in SAS to calculate adjusted FDR p-values, and evaluated those p-values against an alpha of .05 (Benjamini & Hochberg, 1995), which resulted in significant p values for two of the outcomes. These were significant group by time effects for the TOWRE-SWE, $F(4, 600) = 5.28$, $p = .0003$, FDR adjusted $p = .0013$ and the WJ-III Letter/Word ID, $F(4, 600) = 3.84$, $p = .0018$, FDR adjusted $p = .0046$. All other tests were non-significant.

Next, we conducted additional univariate repeated-measures ANOVAs to test pairwise differences of the linear and quadratic slopes between groups. This resulted in 12 single degree of freedom comparisons, three each for the linear and quadratic slopes for the TOWRE-SWE and WJ-III Letter/Word ID outcomes. We once again used the MULTTEST procedure in SAS to calculate adjusted FDR p-values, and evaluated those p-values against an alpha of .05. As a result, only four of the 12 comparisons were determined to be statistically significant. To aid in the interpretation of these tests, mean values are plotted at each testing point by group for TOWRE-SWE in Figure 1 and WJ-III Letter/Word ID in Figure 2. For the TOWRE-SWE, there was a significant difference in linear slopes between the BAU and the two-year group ($F = 11.35$, $p = .0009$, FDR adjusted $p = .01$), with the two-year group having the greater linear slope. There were also significant differences for the quadratic effect between BAU and the one-year group ($F = 9.96$, $p = .0019$, FDR adjusted $p = .01$). For the WJ-III Letter/Word ID, there were significant differences in linear slope between the BAU and one-year groups ($F = 6.51$, $p = .0115$, FDR adjusted $p = .01$) as well as the BAU and two-year groups ($F = 9.39$, $p = .0025$, FDR adjusted $p = .03$), indicating that both treatment conditions had greater linear slopes than the BAU group. No other comparisons met the critical level of alpha.

DISCUSSION

As hypothesized, students assigned to two years of researcher-provided intervention made statistically significant gains in word reading and fluency when compared to students who received one year of intervention or students assigned to the BAU comparison condition. However, this differential growth in word-level reading was not associated with corresponding differential gains in reading comprehension; we found no statistically significant differences between groups of students randomized to one or two years of researcher-provided intervention or the BAU comparison condition on standardized measures of reading comprehension.

TABLE 3
Tests of Group by Time Interactions and Slope Effects for Statistically Significant Outcomes

Test	*Outcome*			*F*	*p*	*p (FDR-adjusted)*
Group by Time Interaction						
	TOWRE Sight Word Efficiency			5.45	.0003	.0013*
	WJ-III Letter Word ID			4.33	.0018	.0046*
	WJ-III Spelling			1.62	.1677	.2796
	Gates-MacGinitie			0.45	.7745	.7776
	WJ-III Passage Comprehension			0.44	.7776	.7776
Test of Slope Effects		*Slope*	*Comparison*			
	TOWRE Sight Word Efficiency					
		Linear	1YR vs. BAU	3.94	.0487	.0731
		Linear	2YR vs. 1YR	2.55	.1116	.1488
		Linear	2YR vs. BAU	11.35	.0009	.01*
		Quadratic	1YR vs. BAU	9.96	.0019	.01*
		Quadratic	2YR vs. 1YR	4.72	.0309	.0585
		Quadratic	2YR vs. BAU	0.72	.3987	.4784
	WJ-III Letter Word ID					
		Linear	1YR vs. BAU	6.51	.0115	.0345*
		Linear	2YR vs. 1YR	0.6	.4404	.4805
		Linear	2YR vs. BAU	9.39	.0025	.01*
		Quadratic	1YR vs. BAU	5.18	.0239	.0574
		Quadratic	2YR vs. 1YR	4.55	.0341	.0585
		Quadratic	2YR vs. BAU	0.03	.8519	.8519

WJ-III LWID = Woodcock Johnson-Third Edition: Letter Word Identification. TOWRE = Test of Word Reading Efficiency; WJ-III PC = Woodcock Johnson-Third Edition: Passage Comprehension.

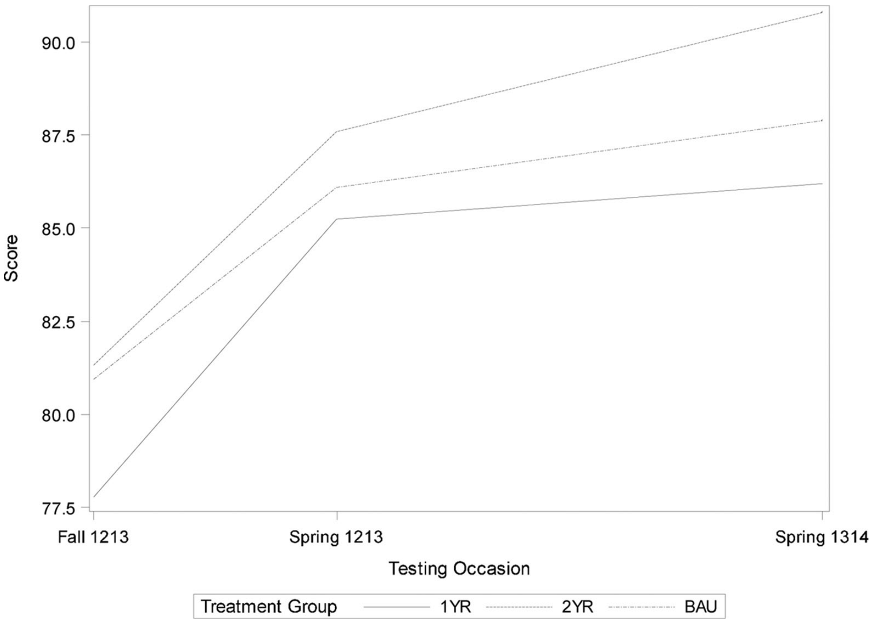

FIGURE 1 Plot of TOWRE Sight Word Efficiency Subtest means by group over time.

Word Reading and Reading Comprehension

For word reading and fluency outcomes, students who received two years of the researcher-provided intervention grew more than both the one-year and BAU groups on both the WJ-III Letter/Word Identification and TOWRE Sight Word Efficiency subtests. Additionally, there were statistically significant differences in slopes for the WJ-III Letter/Word Identification subtest for students in the one-year group and the BAU group, with the one-year group growing at an accelerated rate and closing the gap with the BAU group by the end of fifth grade. At a surface level, these findings likely reflect the larger relative importance assigned to teaching word-level processes within the

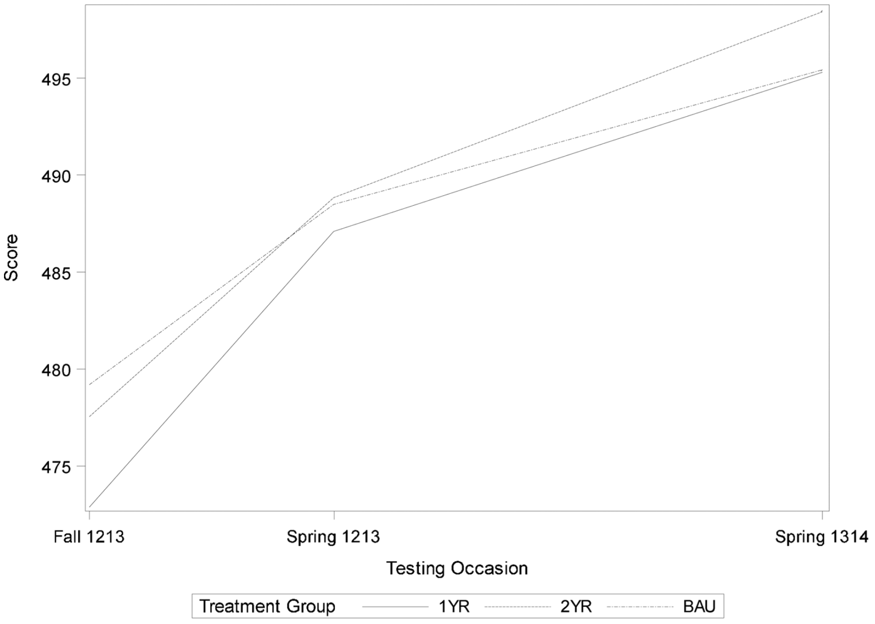

FIGURE 2 Plot of WJ-III Letter Word Identification Subtest means by group over time.

researcher-provided intervention when compared to school-provided interventions. Anecdotal evidence collected as part of the ARI instrument suggested that school-provided interventions were more likely to include text-level activities, most frequently in the form of preparation for high-stakes assessments, guided reading activities, or review of previously taught texts. Few schools reported daily, systematic instruction in word study as was included in the researcher-provided intervention.

However, the differences observed in slopes between the three treatment groups on word level outcomes were not observed on standardized measures of reading comprehension, the primary outcome of interest for this study. This finding is consistent with previous experimental research with struggling readers in late elementary and secondary grades, which generally display relatively larger effects on word level and fluency outcomes than on standardized measures of reading comprehension (Wanzek et al., 2010).

Interpreting Reading Comprehension Results

To interpret the lack of statistically significant results for reading comprehension in the present study, it is important to state the null hypothesis: that differences in slope on standardized measures of reading comprehension are not *detectably larger* for students assigned to the two-year group and one-year group in comparison to the BAU comparison condition (Seftor, 2016). In the sections that follow, we describe the results of this study as yielding few statistically significant between-groups differences and avoiding terms like no effects, null effects, or ineffective; the effectiveness (in absolute terms) of this intervention cannot be inferred from the present study. Rather, this study presents the results of a relative comparison between a researcher-provided intervention (in different doses) and established school curricula and intervention programs—instantiated as the BAU comparison condition. Across all three groups, students spent significant amounts of time in reading interventions, which were differentiated here by the amount (dose) and relative time in school-provided versus researcher-provided interventions. Thus, the inference that these data permit is that assignment to the researcher-provided intervention did not yield detectably large (positive or negative) enough effects on reading comprehension to be differentiated from the BAU comparison group. Why?

The Dynamic Counterfactual

One potential explanation that we have offered previously (Vaughn et al., 2016) posits that academic progress within the counterfactual condition, herein instantiated as a BAU comparison condition, may confound comparisons between the BAU and researcher-provided treatment groups (Lemons, Fuchs, Gilbert, & Fuchs, 2014). We deem the counterfactual as "dynamic" due to the large differences in treatments provided to students and the continuous shifts of interventions provided by the district based on their own analysis of student-level data. This hypothesis is bolstered by an inspection of norm-based standard scores; regardless of condition, participants in all three groups made standard score gains in the two years of the study, especially on the Gates-MacGinitie reading comprehension test. Additionally, it is important to note again that a large number of students across all three groups—including the BAU comparison group—received some form of supplemental reading intervention(s). For example, across both years the number of mean hours spent in school-based reading interventions was over 50 hours, which is greater than the number of mean hours spent in the researcher-provided intervention in either year of the study. This greatly diminishes the dosage contrast between groups, in which the two-year group received only 40 hours of intervention more than the BAU comparison group. Thus, it may be fair to assert that the lack of detectable differences between the groups does not necessarily suggest that the researcher-provided treatment was ineffective. Instead, it is possible that factors present across all groups, including effective Tier 1 instruction and the provision of small-group reading interventions, resulted in reading comprehension gains for all students regardless of condition.

Given the magnitude of initial reading deficits of the students in this study, the improvements in standard scores observed for all groups after one and two years are noteworthy. Few prior studies of reading intervention for students with reading difficulties beyond the primary grades have focused on students with this magnitude of reading delay, and fewer still have used a rigorous RCT design with large samples. For example, Frijters et al. (2013) conducted a reading intervention with middle school students whose pretest standard scores on WJ-III subtests ranged from 76 to 82 standard score points. Similar to the study reported here, the treatment students made substantial standard score gains. However, in Frijters et al., these gains were not matched by students in the comparison condition. Thus, significant differences between groups emerged.

Thus, we interpret the findings relative to the intervention practices described herein as requiring additional evaluation, and consider that interventions for students with significant reading deficits may require even more intensive treatments (e.g., small groups or more customized). However, we must temper this optimism with a frank admission of the limitations of this assertion. There are many potential explanations for standard score gains observed in the present study, reasons both instructional and statistical (i.e., regression to the mean). The present design cannot rigorously test these explanations—an inherent limitation of educational research in which a blinded, no-treatment control condition is not available.

How Malleable are Reading Skills at this Age Among Students with this Degree of Difficulty?

A second potential explanation for the lack of differential results for reading comprehension outcomes points to the diminished malleability of reading comprehension in comparison to word-level reading skills in late elementary grades and beyond. This interpretation disregards the observed standard score gains as artifacts of measurement error or regression to the mean, and posits that the failure to find between groups differences is because the targeted construct (reading comprehension) demonstrates limited malleability and it is unrealistic to expect robust growth in interventions dosages that are measured in hours rather than years. Standardized mean differences for annual growth in reading achievement diminish dramatically as students progress from early elementary grades into late elementary and beyond (Lipsey et al., 2012; Scammacca, Fall, & Roberts, 2015aa). This phenomenon is most likely due to changes in the reading task over time. In earlier grades, text complexity is limited; constrained word-level skills are highly predictive of reading comprehension (Schulte et al., 2016). In later grades, the broader constructs of language and background knowledge are increasingly predictive of reading comprehension (Ahmed et al., 2016; Catts, Adlof, & Weismer, 2006). Thus, to improve reading comprehension in late elementary and secondary grades, it may be necessary to affect the relatively unconstrained constructs of language and background knowledge—a difficult task in under 40 hours of intervention per year. Perhaps the results of this study and others that fail to find robust differential effects in reading comprehension indicate that the task of remediating persistent reading comprehension deficits in late elementary and secondary school will require interventions of greater duration and dosage than previously studied.

It is also possible to interpret our inability to find robust between-group differences as a failure of the underlying theory of change, premised on the SVR. Although the SVR has proven a particularly valuable heuristic for the partitioning of variance in reading comprehension, it may not provide sufficient guidance for the development of robust interventions, particularly in later grades (Compton et al., 2014). The complexities of the linguistic comprehension component need to be further considered and investigated with an eye towards providing better direction on what particular components of language have the most robust impact on understanding of text, which could be further considered as guidance for intervention components.

Why is this Study Difficult to Publish?

The theme of this special issue is studies with null results and the difficulties associated with publishing such studies. This difficulty is a problem if it results in publication bias, which is shorthand for a complex socio-scientific problem in which researchers and journals are reluctant to publish studies that do not demonstrate statistically significant results, often independent of the rigor of the research design (Cook & Therrien, 2017; Greenwald, 1975; Ioannidis, 2005). This bias is most acutely observed in meta-analysis, in which mean effect sizes may be systematically inflated due to an absence of published null or negative-results studies (Ferguson & Brannick, 2012; Gage, Cook, & Reichow, 2017; Pigott, Polanin, Valentine, Williams, & Canada, 2013). However, publication bias may also create false impressions among researchers and practitioners about the relative malleability of specific psycho-educational domains, and about what constitutes effective treatment.

Polanin, Tanner-Smith, and Hennessy (2016) recently completed a meta-analysis of meta-analyses in psychological and educational research to evaluate to what extent publication bias is observed. Effect sizes from published and unpublished literature (so-called "gray literature") were compared utilizing data from 81 meta-analyses published between 1986 and 2013. The results find strong evidence for publication bias: published studies demonstrated much larger mean meta-analytic effect sizes (+0.18) than unpublished studies—a standardized mean difference nearly equal to the mean meta-analytic effect size for intensive interventions in late elementary and secondary grades (+0.21; Scammacca et al., 2015a). Gage et al. (2017) completed a similar study of meta-analyses published in special education journals. Mirroring Polanin et al. results indicated that meta-analyses that did not include gray literature were more likely to demonstrate publication bias, and published studies yielded effect sizes of greater magnitude than unpublished studies. Such results are troubling, and have important implications for the dissemination and interpretation of educational research. Publication bias represents a multi-faceted problem, with culpability falling upon reviewers, journal editors, and authors – though summaries of problems and potential solutions for reviewers and journal editors have been written (e.g., Miguel et al., 2014). In the sections that follow, we highlight the difficulties that have confronted our research group in publishing high-quality studies that demonstrate few statistically significant results.

Complicated Interpretation

The tenuous language throughout the discussion of the present study hints at one significant challenge to the publication of studies that fail to find statistically significant differences between groups: it is often difficult to determine exactly what happened. In contrast to laboratory settings, where control conditions are observed or are equivalent to no treatment, when an educational intervention does not outperform the normative treatment, there are several broad explanations for why no differences emerged: (1) the theory of change failed, (2) the implementation of the intervention failed, or (3) the research design failed (Seftor, 2016). Importantly, the results alone cannot point to which reason or combination of reasons is most pertinent. Instead, as researchers we must parse each of these possibilities and evaluate its likelihood—an uncertain and uncomfortable exercise. In the present study, we identified two potential explanations for the observed pattern of results, one pointing to a failure of our theory of change (i.e., reading comprehension is not sufficiently malleable to respond to this treatment) and one pointing to a failure of the research design (i.e., growth in the BAU confounded our treatment contrast). Based on fidelity of implementation observations, we are reasonably confident that the intervention was implemented as intended. Yet there is no certainty in these explanations, complicating both interpretation and the publication process. This uncertainty may represent a partial explanation for why authors are less likely to submit for publication papers that lack statistically significant results (Cooper, DeNeve, & Charlton, 1997).

Statistical Significance and Effect sizes

Fletcher and Wagner (2014) have argued that intervention research should de-emphasize statistical significance and *p* values and instead focus on the importance of small but meaningful effect sizes that can accumulate over time. For example, in a previous intervention study in which students in both the researcher-provided intervention and the BAU comparison condition demonstrated robust growth but there were few statistically significant differences between groups, we interpreted the effects of the intervention as impactful because growth in both groups exceeded normative expectations and there were trends toward the researcher-provided treatment (Vaughn et al., 2016). This reframing of how researchers and practitioners might interpret intervention effects has considerable merit, but effect sizes are not a panacea; there are many study designs in which a simple comparison of between-groups effect sizes at post-test may not aid in interpretation. For example, in the present study we reported relatively large differences between groups at pretest, despite random assignment (see Table 2; *d* range at pretest: /0.03/-/0.40/). Over the course of the study, these group differences diverge, dependent upon the measure. To account for this fan-shaped growth and evaluate change, we chose to evaluate slope using all three assessment time points, with *post hoc* tests comparing univariate group slopes over time. Such a design does not lend itself well to simple interpretation of group mean differences represented by effect sizes. Thus, for a complex study like this one, the most readily interpretable metric by which to evaluate the null hypothesis may be *p*-values, leaving less interpretative room than a relatively simpler study with two similar groups and straightforward effect sizes.

Complicated Review Process

Another explanation for authors being reluctant to submit studies that do not demonstrate statistically significant results is that the peer-review process is longer and more complicated (Suñé, Suñé, & Montoro, 2013). This explanation is consistent with our own experience, in which we have observed that peer review for studies with few or no statistically significant results requires additional requests for analyses, often unnecessarily creating additional work that adds little to the interpretation of the results. This observation is not a wholesale critique of the peer-review process. However, in recent publications demonstrating few statistically significant results, we have endured lengthy and complicated reviews. In our experience, these complicated reviews are not due to explicit bias against publishing studies without statistically significant findings; we are optimistic that most journals and reviewers are increasingly aware that the dissemination of well-designed studies is important regardless of the specific findings. Instead, we have observed a contrary tendency in which earnest reviewers want to help us find statistically significant results. This wish to help often consists of voluminous questions about the many design and analytic decisions we made throughout the study. These questions can necessitate long review responses and frequently require additional analyses and revisions. In contrast, studies that demonstrate statistically significant results do not foster the same tendency.

Implications for Practice and Research

Science is a cumulative activity, and no single study should be used as the basis for high-stakes decision making. This study should be interpreted in the context of other studies and reviews that similarly report small effect sizes for reading interventions targeting reading comprehension in late elementary and secondary settings (Scammacca et al., 2015a; Wanzek et al., 2010). This growing body of literature affirms one key takeaway: there is no silver bullet to remediate years of difficulty in reading. With this reality in mind, practitioners should work to avoid isolated, piecemeal intervention strategies. Successful intervention programs will require high-quality, long term interventions in which effects cumulate across years (Fletcher & Wagner, 2014). Similarly, researchers should redouble efforts to maximize the effects of interventions, to identify causal mechanisms, and to study interventions of greater intensity and duration than are typically studied.

CONCLUSION

We evaluated the effects of a researcher-provided intervention, provided for two years or one year, compared to a BAU comparison condition. There were statistically significant differences for students in the two-year group compared to the BAU group on measures of word reading and fluency. However, we found no statistically significant differences between groups on standardized measures of reading comprehension. We discussed these findings in the context of inherent limitations of school-based research, highlighting the importance of relative comparisons and the questionable malleability of reading comprehension at this age.

Acknowledgments

This research was supported by Award Number P50 HD052117, Texas Center for Learning Disabilities, from the Eunice Kennedy Shriver National Institute of Child Health & Human Development to the University of Houston. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health & Human Development or the National Institutes of Health.

REFERENCES

Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, *44*, 68–82.

Beck, I. L., & McKeown, M. G. (2006). *Improving comprehension with questioning the author: A fresh and enhanced view of a proven approach*. New York: Scholastic.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Murray, M. S., Munger, K. A., & Vaughn, M. G. (2014). Intensive reading remediation in grade 2 or 3: Are there effects a decade later? *Journal of Educational Psychology*, *106*, 46–57.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research*, *49*(2), 278–293.

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*, 55–73.

Cook, B. G., & Therrien, W. J. (2017). Null effects and publication bias in special education research. *Behavioral Disorders*, *42*, 149–158.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*, 447–452.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120–128.

Fletcher, J. M., & Wagner, R. K. (2014). Accumulating knowledge: When are reading intervention results meaningful? *Journal of Research on Educational Effectiveness*, *7*, 294–299.

Flynn, L. J., Zheng, X., & Swanson, H. L. (2012). Instructing struggling older readers: A selective meta-analysis of intervention research. *Learning Disabilities Research & Practice*, *27*, 21–32.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, *88*(1), 3–17.

Frijters, J. C., Lovett, M. W., Sevcik, R. A., & Morris, R. D. (2013). Four methods of identifying change in the context of a multiple component reading intervention for struggling middle school readers. *Reading and Writing*, *26*, 539–563.

Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, *83*, 428–445.

Gersten, R., Baker, S. K., Smith-Johnson, J., Dimino, J., & Peterson, A. (2006). Eyes on the prize: Teaching complex historical content to middle school students with learning disabilities. *Exceptional Children*, *72*, 264–280.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*, 6–10.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Gwet, K. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD: STATAXIS Publishing Company.

Hiebert, E. H. (2003). *QuickReads – A research-based fluency program*. Parsippany, NJ: Pearson.

Hirsch, E. D. (2006). Reading-comprehension skills? What are they really? *Education Week*, *25*, 52. Retrieved from https://du.idm.oclc.org/login?url=https://search-proquest-com.du.idm.oclc.org/docview/202726235?accountid=14608.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*, 127–160. https://doi.org/10.1007/BF00401799

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

James-Burdumy, S., Deke, J., Gersten, R., Lugo-Gil, J., Newman-Gonchar, R., Dimino, J., et al. (2012). Effectiveness of four supplemental reading comprehension interventions. *Journal of Research on Educational Effectiveness*, *5*, 345–383.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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*, 242–252.

Lesnick, J., Goerge, R., Smithgall, C., & Gwynne J. (2010). *Reading on grade level in third grade: How is it related to high school performance and college enrollment*? Chicago: Chapin Hall at the University of Chicago.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., et al. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. (NCSER 2013–3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

MacGinitie, W., MacGinitie, R., Maria, K., Dreyer, L., & Hughes, K. (2000). *Gates-MacGinitie Reading Tests-4*. Itasca, IL: Riverside.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). Technical manual. *Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside Publishing.

McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, *44*, 218–253.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., et al. (2014). Promoting transparency in social science research. *Science*, *343*, 30–31.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*, 207–236.

Pigott, T. D., Polanin, J. R., Valentine, J. C., Williams, R. T., & Canada, D. D. (2013) Outcome-reporting bias in education research. *Educational Researcher*, *42*, 424–432.

Scammacca, N. K., Fall, A. M., & Roberts, G. (2015a). Benchmarks for expected annual academic growth for students in the bottom quartile of the normative distribution. *Journal of Research on Educational Effectiveness*, *8*(3), 366–379.

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

Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. T. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test? *Journal of Educational Psychology*, *108*, 925–942.

Seftor, N. (2016). *What does it mean when a study finds no effects*? Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Solis, M., Ciullo, S., Vaughn, S., Pyle, N., Hassaram, B., & Leroux, A. (2012). Reading comprehension interventions for middle school students with learning disabilities: A synthesis of 30 years of research. *Journal of Learning Disabilities*, *45*, 327–340.

Suñé, P., Suñé, J. M., & Montoro, J. B. (2013). Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. *PLoS One*, *8*, e54583.

Torgesen, J., Wagner, R., & Rashotte, C. (1999). *Test of word reading efficiency*. Austin, TX: Pro-Ed.

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1998). The landscape model of reading: Inferences and the on-line construction of a memory representation. In H. Van Oosendorp & S.R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71–98). Mahwah, NJ: Erlbaum.

Vaughn, S., Solís, M., Miciak, J., Taylor, W. P., & Fletcher, J. M. (2016). Effects from a randomized control trial comparing researcher and school-implemented treatments with fourth graders with significant reading difficulties. *Journal of Research on Educational Effectiveness*, *9*, 23–44.

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.

Wolff, U. (2016). Effects of a randomized reading intervention study aimed at 9-year-olds: A 5-year follow-up. *Dyslexia*, *22*, 85–100.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Rolling Meadows, IL: Riverside Publishing.

About the Authors

Dr. Jeremy Miciak is an Assistant Research Professor in the Department of Psychology at the University of Houston and a faculty member of the Texas Institute for Measurement, Evaluation, and Statistics. His research focuses on students with or at-risk for learning disabilities.

Dr. Garrett Roberts is a board-certified behavior analyst at the doctoral level an Assistant Professor in the Teaching and Learning Sciences department at The University of Denver. His research focuses on developing and implementing reading and behavioral interventions with an emphasis on learning disabilities and associated deficits in attention and behavior.

Dr. Pat Taylor is an Assistant Research Professor at the University of Houston and a faculty member of the Texas Institute for Measurement, Evaluation, and Statistics. His research focuses on the development and application of advanced statistical methods in educational research, particularly in studies evaluating students with learning difficulties.

Dr. Michael Solis is an Assistant Professor of special education at the University of California Riverside. His line of research focuses on vocabulary and reading comprehension interventions for students with reading difficulties in grades 4–12. He has expertise in reading interventions, multi-tiered systems of support, behavior interventions, and collaboration.

Dr. Yusra Ahmed is a Research Assistant Professor of Developmental, Cognitive and Behavioral Neuroscience in the Psychology Department at the University of Houston, and a faculty member of the Texas Institute for Measurement, Evaluation and Statistics. Her current research interests include the co-development of reading and writing in relation to typical development and in children with learning disabilities.

Dr. Sharon Vaughn, Manuel J. Justiz Endowed Chair in Education, is the Executive Director of The Meadows Center for Preventing Educational Risk an organized research unit at The University of Texas at Austin. She is currently Principal Investigator on several Institute for Education Sciences, National Institute for Child Health and Human Development, and U.S. Department of Education research grants.

Dr. Jack M. Fletcher is a Hugh Roy and Lillie Cranz Cullen Distinguished Professor of Psychology at the University of Houston. For the past 30 years, Dr. Fletcher, a board-certified child neuropsychologist, has worked on issues related to child neuropsychology, including studies of children with spina bifida, traumatic brain injury, and other acquired disorders. In the area of developmental learning and attention disorders, Dr. Fletcher has addressed issues related to definition and classification, neurobiological correlates, and most recently, intervention.