THE RELATIVE IMPACT OF ALIGNING TIER 2 INTERVENTION MATERIALS WITH CLASSROOM CORE READING MATERIALS IN GRADES K–2

ABSTRACT

This randomized controlled trial in 55 low-performing schools across Florida compared 2 early literacy interventions—1 using stand-alone materials and 1 using materials embedded in the existing core reading/language arts program. A total of 3,447 students who were below the 30th percentile in vocabulary and reading-related skills participated in the study. Both interventions were implemented with fidelity for 45 minutes daily for 27 weeks in small groups of 4 students (or 5 in grade 2). The stand-alone intervention significantly improved grade 2 spelling outcomes relative to the embedded intervention; there were some differential impacts due to cohort and baseline and, in kindergarten, to English-learner status. On average, students in schools in both interventions showed similar improvement in reading and language outcomes and similar percentile gains to those in recent systematic reviews. Results are discussed with respect to alignment of Tier 2 instruction with Tier 1 instruction.

Barbara R. Foorman
Sarah Herrera
Jennifer Dombek
FLORIDA STATE UNIVERSITY

THE ELEMENTARY SCHOOL JOURNAL
Volume 118, Number 3. Published online January 17, 2018

THERE is a strong research base on the skills needed for effective early reading intervention (e.g., Foorman, Beyler, et al., 2016). Components of effective early reading intervention include (*a*) explicit instruction in phonological awareness, linking letters to sounds, decoding, and word study, and (*b*) practice reading text for accuracy, fluency, and comprehension (e.g., Foorman, Beyler, et al., 2016; Foorman & Connor, 2011; National Institute of Child Health and Human Development, 2000; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). Recent research has demonstrated the efficacy of directly teaching academic language to students to improve their comprehension (Baker et al., 2014; Catts, Nielsen, Bridges, & Liu, 2014; Foorman, Beyler, et al., 2016). Academic language is "the formal communication structure and words that are common in books and at school" (Foorman, Beyler, et al., 2016, p. 7). These reading and language skills—that is, literacy skills—are often delivered in multiple tiers of instruction that include the classroom at Tier 1; supplementary, small-group intervention at Tier 2; and intensive, small-group intervention at Tier 3 for students not making progress with Tier 2 intervention (e.g., Fletcher & Vaughn, 2009; Gersten et al., 2009; Gersten, Jayanthi, & Dimino, 2017). This cluster randomized study in 55 schools across Florida examines the relative impact of aligning Tier 2 instructional materials with the Tier 1 materials adopted to represent the Florida standards (which are essentially English language arts Common Core standards plus cursive writing).

Effective Tier 2 Literacy Interventions in the Primary Grades

Two recent reviews of the reading intervention literature, using What Works Clearinghouse (WWC; WWC, 2014) design standards, provide a list of effective early reading interventions. Key characteristics of studies that meet these design standards are randomized controlled trials (RCTs) or quasi-experimental design studies that establish baseline equivalence.

Foundational Reading Skills Practice Guide

The first systematic review was of the practice guide, *Foundational Skills to Support Reading for Understanding in Kindergarten through 3rd Grade* (Foorman, Beyler, et al., 2016). This guide considered outcomes in 12 domains related to early reading skills: encoding, general achievement, letter names and sounds, listening comprehension, morphology, oral reading accuracy, oral reading fluency, phonology, reading comprehension, syntax, vocabulary, and word reading. Eligibility criteria included studies of K–3 children participating in general education classes and excluded samples in which more than 50% of children had identified disabilities or were English learners. Of the 4,500 studies screened that were published from 2000 through 2014, 56 met WWC design standards, and recommendations for practice are based on those 56 studies (characteristics of the 56 studies are described in App. D in the practice guide). The majority of the 56 studies were Tier 2 one-on-one or small-group interventions delivered by tutors, paraprofessionals, or graduate assistants to at-risk students. The four recommendations and their levels of evidence are (*a*) teach students academic language skills, including the use of

inferential and narrative language and vocabulary knowledge (minimal evidence); (*b*) develop awareness of the segments of sounds in speech and how they link to letters (strong evidence); (*c*) teach students to decode words, analyze word parts, and write and recognize words (strong evidence); and (*d*) ensure that each student reads connected text every day to support reading accuracy, fluency, and comprehension (moderate evidence).

Systematic Review of Response to Intervention Reading by the Regional Educational Laboratory Southeast

The second systematic review was conducted by Gersten, Newman-Gonchar, Haymond, and Dimino (2017) for the Regional Educational Laboratory (REL) Southeast. They assessed the evidence base of research supporting reading interventions in grades 1–3 in literature published from January 2002 to June 2014. They restricted their search to studies with RCT or quasi-experimental designs, samples of at-risk students in grades 1–3 scoring below (or predicted to be below) the thirty-fifth percentile on a valid and reliable screener, reading interventions that lasted more than 8 hours, and studies that included at least one reading outcome.

Of the 1,813 studies screened, 43 met screening criteria, and 23 of those met WWC standards. Instruction in these 23 studies primarily focused on phonemic awareness, phonics, and reading at the word and sentence levels in grade 1 and on decoding, encoding (spelling), vocabulary, and comprehension in grades 2 and 3. Interventionists were well trained, monitored for fidelity, and provided with ongoing support. Group size in the 23 studies tended to be more one on one than small group, but differential effects based on group size were minimal. However, the average size of the small groups was relatively small: 3.4 students in grade 1 and 2.8 students for grades 2 and 3. This approach contrasts with the larger average group sizes reported by Balu et al. (2015) in a national evaluation of response to intervention (RtI) in elementary schools based on a regression discontinuity design: 5.3 students in grade 1, 5.9 in grade 2, and 6.4 in grade 3. Gersten, Jayanthi, et al. (2017) suggest that the nonexistent and even negative effects for RtI found by Balu et al. (2015) may have been due to the larger group sizes.

Gersten and colleagues (Gersten, Jayanthi, et al., 2017; Gersten, Newman-Gonchar, et al. 2017) note that their systematic review yielded findings similar to the meta-analysis of K–3 reading interventions conducted by Wanzek et al. (2016). Effects were strongest for pseudoword and real-word decoding, with typical gains of 13 to 17 percentile points. Effects for reading comprehension were approximately 13 percentile points, on average.

Alignment of Tier 2 Literacy Intervention with Tier 1 Classroom Instruction

Fien and colleagues (Baker, Fien, & Baker, 2010; Fien et al., 2015; Smith, Smolkowski, Baker, Fien, & Kosty, 2016) argue that multitiered interventions are most effective when they consist of explicit Tier 1 reading instruction and Tier 2 intervention that is aligned with the scope and sequence of Tier 1 instruction. They designed a

multitiered classroom instruction and intervention model called "enhanced core reading instruction" (ECRI). Fien et al. (2015) tested the efficacy of ECRI in grade 1 by randomly assigning 16 schools to ECRI or to a control. Every day, at-risk students (i.e., those below the thirty-first percentile on the Stanford Achievement Test [SAT]; $n = 267$) received 90 minutes of explicit, whole-group classroom instruction and 30 additional minutes of aligned, small-group Tier 2 intervention. The intervention showed statistically significant positive effects on students' decoding and first-semester reading fluency and potentially positive effects on reading comprehension and total reading achievement. Similar results were obtained by Smith et al. (2016) in their cluster RCT of ECRI in grade 1 in 44 elementary schools, blocked by district. Significant treatment effects were found on measures of phonemic decoding and oral reading fluency from fall to winter and on word reading from fall to spring. Analysis of classroom observation data indicated that students in ECRI treatment schools made significantly greater gains than those in comparison schools when the quality of explicit instruction was high.

Selection of the Tier 2 intervention materials embedded within the core reading/language arts program is an attractive option for schools because these materials are aligned with standards and classroom instruction and do not require the purchase of additional materials. However, as research based as these Tier 2 materials embedded within the core reading/language arts program may claim to be, they are rarely evaluated empirically. Therefore, another option for schools is to select Tier 2 materials and strategies outside of the core reading/language arts program that have credible evidence of effectiveness on reading and language outcomes, such as the ratings provided by the WWC. Given the evidence base, it is plausible to expect relatively better small-group, Tier 2 intervention outcomes in reading and language in schools randomly assigned to stand-alone materials than to schools randomly assigned to embedded materials. However, the use of embedded Tier 2 materials may prove relatively more effective because of their alignment with the core reading/language arts program and thus with instructional practice.

Purpose

This study was conducted by the REL Southeast in partnership with its Improving Literacy Research Alliance, which was a group of state and local education agency practitioners interested in improving literacy outcomes in low-performing public schools. Alliance members were particularly interested in studying effective ways of implementing Tier 2, small-group literacy intervention in grades K–2 in both urban and rural schools, so as to reduce the numbers of students not meeting grade-level proficiency standards in grade 3 (Foorman, Herrera, Dombek, Schatschneider, & Petscher, 2017). Moreover, alliance members thought that the academic language component added to the reading component of the embedded and stand-alone interventions would be helpful to all at-risk students in low-performing schools, but particularly to English-learner students.

This study was conducted across 2 school years—2013–2014 and 2014–2015—in 55 Florida elementary schools, randomly assigned within south, central, and northern regions, to 1 year of embedded or stand-alone intervention. The intervention

was implemented in small groups of four students in kindergarten and grade 1 and five students in grade 2 for 45 minutes daily for a total of 27 weeks. Alternative designs could have entertained a no-intervention or business-as-usual control. However, this study was conducted in Florida, where districts encourage intervention for at-risk students. In addition, business as usual in Florida schools typically means 30 minutes of reading intervention, not the 45 minutes of reading and language intervention implemented in this study. Finally, "business as usual" means something different from school to school and, even within schools, and is constantly changing over time (Lemons, Fuchs, Gilbert, & Fuchs, 2014).

The following research question was analyzed separately by grade: What is the impact of a stand-alone early literacy intervention relative to an embedded early literacy intervention on reading and language outcomes? Within this overarching question were two subquestions: (*a*) Is the impact different between the two cohorts of schools or dependent on baseline scores? (*b*) Are there differences in reading and language outcomes between stand-alone and embedded interventions for English-learner and non-English-learner students? Are there differences within interventions between English-learner and non-English-learner students?

Method

Participants

Schools. To ensure that the embedded intervention meant the same thing across districts, five districts from across Florida that had selected the most widely adopted core reading/language arts program in Florida for their elementary schools—Houghton Mifflin Harcourt (HMH) *Journeys*—were invited and subsequently agreed to participate and to have their participating schools randomly assigned to embedded or stand-alone intervention. Altogether, 55 unique schools participated over 2 school years: 2013–2014 and 2014–2015 (hereafter referred to as Cohort 1 and Cohort 2, respectively; see Fig. 1). Twenty-seven schools participated in Cohort 1 and 28 schools participated in Cohort 2. These schools were in districts that represented the geographic diversity of Florida: one large urban district in south Florida (with 16 different participating schools in each cohort); one medium urban district in central Florida (with eight participating schools in Cohort 1 and nine different schools in Cohort 2); and three small, rural districts in north Florida (two rural districts in Cohort 1 with three schools, and one rural district in Cohort 2 with three schools).

Districts targeted their lowest performing schools for participation—those with school grades of C, D, and F (calculated based on the percentage of students who were proficient and the percentage making learning gains on the state reading test). REL Southeast staff members randomly assigned schools within region and cohort to stand-alone or embedded intervention, according to the following steps: (*a*) assign a random number to each school by region within cohort, (*b*) order schools in descending order within each region and cohort by the assigned random number, and (*c*) assign the first half within region and cohort to the stand-alone intervention group and the second half to the embedded intervention group. Across all schools, 74% of the kindergarten students (75% in Cohort 1 and 73% in Cohort 2), 75% of the grade 1 students (73% in Cohort 1 and 77% in Cohort 2), and 76%

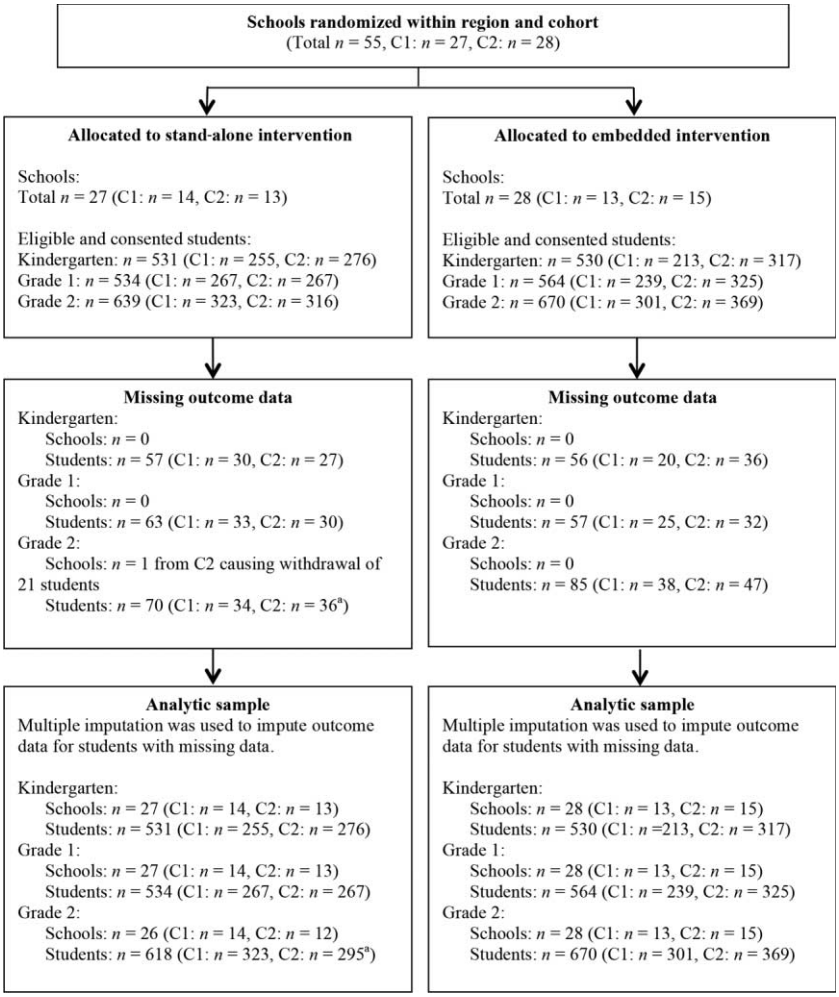

Figure 1. Student and school CONSORT (Consolidated Standards of Reporting Trials) diagram for grades K–2. C1 = Cohort 1; C2 = Cohort 2. [a]Does not include the 21 students attending the Cohort 2 school that withdrew from the study because of scheduling conflicts.

of the grade 2 students (74% in Cohort 1 and 77% in Cohort 2) qualified for free or reduced-price lunch.

Students. In late September, classroom teachers sent parental consent forms home with students who performed below the thirtieth percentile on one or more of the following K–2 tasks from the Florida Center for Reading Research Reading Assessment (FRA; Foorman, Petscher, & Schatschneider, 2015): phonological awareness (kindergarten only), word reading (grades 1 and 2), and vocabulary pairs (in all three grades). Students who were already receiving school services (e.g., special education) were removed from the list of eligible students. In late September, school staff members examined students' schedules to determine which of the remaining eligible students could be served in the daily 45-minute periods available in the bell schedule for small-group intervention and sent home parental consent forms with those students. School staff members continued to send home parental consent forms with

additional students who fit both the eligibility and scheduling criteria until the number of participants determined by the statistical power analysis was achieved.

The total number of eligible students with parental consent was 1,061 ($n = 468$ in Cohort 1 and $n = 593$ in Cohort 2) in kindergarten, 1,098 ($n = 506$ in Cohort 1 and $n = 592$ in Cohort 2) in grade 1, and 1,309 ($n = 639$ in Cohort 1 and $n = 670$ in Cohort 2) in grade 2. Students were placed into intervention groups of four to five students within grade but across classrooms based on whether their FRA screening scores on phonological awareness in kindergarten, word reading in grades 1 and 2, or vocabulary in all grades were relatively high, medium, or low. Table 1 contains student demographic information by cohort and grade.

Figure 1 reports school and student sample sizes at randomization by intervention, cohort, and grade, as well as the number of students missing outcome data and information about the analytic sample. One of the participating 28 schools in the stand-alone intervention in Cohort 2 was excluded from the grade 2 analyses because scheduling conflicts resulted in the withdrawal of all participating grade 2 students at that school ($n = 21$). The proportion of students across grades, intervention conditions, and cohorts that did not complete outcome testing ranged from 10.5% to 12%. Student-level differential attrition between the two interventions across grades and cohorts ranged from 0.2% to 2.4%. School-level attrition for kindergarten and grade 1 across cohorts was 0%. However, the loss of a school in grade 2 Cohort 2 resulted in an overall attrition rate of 3.6% and a differential attrition rate of 7.7%. Based on the WWC liberal attrition boundary, school- and student-level attrition rates were considered low for all grades and both cohorts (WWC, 2014).

Interventions

The Tier 2 reading materials embedded within HMH *Journeys* are called "Strategic Intervention." The supplementary language component is called "Curious about Words." Researchers at the REL Southeast and a subcontractor indepen-

Table 1. Student Demographic Information by Intervention, Cohort, and Grade

	Stand-Alone						Embedded					
	Cohort 1		Cohort 2		Full		Cohort 1		Cohort 2		Full	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Kindergarten:												
Male	255	54	276	56	531	55	212	49	317	60	529	55
EL	254	26	258	40	512	33	211	45	289	40	510	42
FRL	254	88	241	88	495	88	212	78	299	88	501	84
Grade 1:												
Male	267	51	267	57	534	54	237	54	325	54	562	54
EL	265	36	258	29	523	33	230	34	314	36	544	35
FRL	265	85	256	82	521	83	230	76	312	85	542	81
Grade 2:												
Male	323	54	316	59	639	56	301	53	369	54	670	54
EL	323	25	308	35	631	30	300	27	347	42	647	35
FRL	323	89	308	81	631	85	300	73	345	86	645	80

Note.—EL = English-learner students; FRL = eligible for free or reduced-price lunch.

dently reviewed the research, including the evidence ratings on the WWC website, for Tier 2 reading interventions that had been studied with at-risk students in grades K–2 and implemented in small groups. The program that met these criteria and had the strongest levels of evidence in alphabetics, fluency, and comprehension was *Sound Partners* (Vadasy & Sanders, 2012; Vadasy, Sanders, & Abbott, 2008; Vadasy, Sanders, & Peyton, 2006). No language intervention programs for at-risk students in grades K–2 had been rated by the WWC. Therefore, a vocabulary program with good clinical evidence called *Bridge of Vocabulary* (Montgomery, 2007) and an inferential language program with evidence of efficacy called *Language in Motion* (Phillips, 2014) were used in addition to *Sound Partners* to create the stand-alone intervention. In a study of *Language in Motion* with 354 at-risk students in kindergarten and grade 1, Phillips (2014) reported effect sizes exceeding .25 on 10 of 16 syntax and listening comprehension outcomes. Components of the stand-alone and embedded interventions are described in Figure 2. Each component is described below.

Stand-alone intervention. The stand-alone intervention consisted of a 25- to 30-minute daily reading component (i.e., *Sound Partners*) and 15-minute daily language component, *Bridge of Vocabulary*, 3 times a week, and *Language in Motion* twice a week.

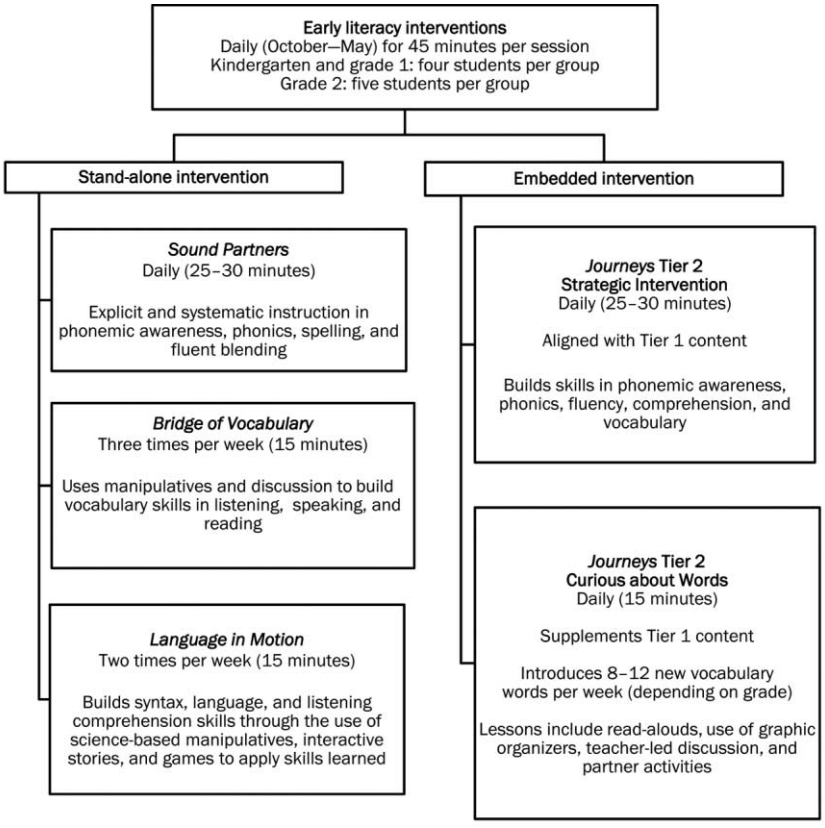

Figure 2. Components of stand-alone and embedded early literacy interventions.

Sound Partners (Vadasy et al., 2004) is a scripted program designed to provide students with explicit and systematic instruction in phonemic awareness, phonics, spelling, and reading, for 30 minutes a day. Daily lessons include decoding and encoding skills taught in isolation and through sentence and storybook reading. Lesson implementation is clearly described in a manual. Kindergarten has its own lesson book, whereas the lessons for grades 1 and 2 are combined into one book. Interventionists are directed to administer skill assessments after every 10 lessons. When skills are not mastered, interventionists must repeat lessons.

Bridge of Vocabulary (Montgomery, 2007) is a semiscripted, explicit vocabulary program that uses manipulatives and discussion to build vocabulary skills. *Bridge of Vocabulary* lessons included in this study focused on building skills in listening, speaking, and reading vocabulary. *Language in Motion* (Phillips, 2014) is a scripted program designed to build syntax, language, and listening comprehension skills through the use of science-based manipulatives, interactive stories, and games.

Embedded intervention. The embedded intervention consisted of a 25- to 30-minute daily reading component (i.e., Strategic Intervention) and 15-minute daily language component (i.e., Curious about Words).

Strategic Intervention is aligned with Tier 1 (i.e., classroom) content and provides students with explicit and systematic instruction in phonemic awareness, phonics, fluency, comprehension, and vocabulary through scripted lessons. Lessons include the use of various word, picture, and vocabulary cards as well as leveled readers and write-in workbooks. In contrast to *Sound Partners*, Strategic Intervention does not include specific provisions for remediation despite including a skills assessment every 10 lessons. Curious about Words, based on Beck, McKeown, and Kucan's (2013) strategies for teaching vocabulary words that are embedded in challenging text read aloud by a teacher, supplements Tier 1 content and introduces students to eight to 12 new vocabulary words each week. Lessons consist of read-alouds, teacher-led discussion, graphic organizers, and partner activities to build students' listening and speaking vocabulary.

As is apparent in Figure 2, both interventions consisted of the foundational components recommended in the literature (e.g., Foorman, Beyler, et al., 2016). The differences were the inclusion of spelling and inferential language in the stand-alone intervention and the inclusion of comprehension in the embedded intervention. An even more critical difference, however, was that, unlike the stand-alone intervention materials, the embedded intervention materials had no implementation manual. Directions for Strategic Intervention were in a tab in the teacher's edition of the core reading program, student anthologies were part of the core reading program, formative assessments were in a separate booklet, and vocabulary cards were in their own box. Therefore, REL Southeast staff developed training materials for the embedded intervention that included information about scope and sequence and instructional procedures.

Procedures

Intervention began in mid-October each year and continued daily in 45-minute sessions until the end of May, for approximately 27 weeks. Students were assessed at baseline in September through early October and at outcome in April through

May. During the first 10 weeks, between 6% and 15% of students across grades K–2 moved to another small group because their scores on skill mastery tests were more similar to the scores of students in another small group.

Interventionists. As an incentive to participate, the REL Southeast hired two to three local interventionists per school who had previously worked with children in educational settings. School leaders were also encouraged to contribute paraprofessionals as interventionists to serve more at-risk students and to build capacity at the school for intervention to continue after the study ended. In Cohort 1, the REL Southeast provided 66 interventionists, and schools provided 17 paraprofessionals; together, they served 370 small groups. In Cohort 2, the REL Southeast provided 64 interventionists, of which 42% had been interventionists the previous year. Schools provided 25 paraprofessionals, and, together with the REL Southeast hires, they served 424 small groups. Interventionists hired by the REL Southeast and paraprofessionals provided by the schools are hereafter collectively referred to as "interventionists." In Cohort 1, 32% of the interventionists were certified teachers, and in Cohort 2, 37% were certified teachers. On average, each school had three to four interventionists, each serving four to six small groups of four to five students across grades K–2.

Professional development and ongoing support. Each year the interventionists were trained over a 2-day period during late September and sent home with the manuals and instructional materials they would be using to familiarize themselves with the strategies, materials, and corresponding skill assessments. During early October, the interventionists visited their assigned school to meet the grade K–2 teachers and school staff and to set up materials in their intervention space.

Once the intervention started in mid-October, the REL Southeast staff members visited each interventionist to answer questions and to offer additional training. A lead interventionist was designated at each school to communicate with school and REL Southeast staff. In addition, interventionists audio-recorded 1 week of lessons each month for periodic review by the REL Southeast staff. The audio recordings were referred to occasionally in discussions of student behavior.

Fidelity of implementation. Fidelity was defined as the percentage of the lesson in which instruction followed the lesson sequence and script within each of the skills taught. The REL Southeast staff observed each intervention group once in the fall and once in the spring and completed separate fidelity checklists for the reading and language components of each intervention. Observers responded "yes," "no," or "not applicable" to the question of whether the lesson sequence and/or script had been followed for each skill. Observers were required to achieve better than 80% reliability on the checklist during training before they were permitted to conduct live observations. Interrater reliability was calculated using Krippendorff's alpha (Hayes & Krippendorff, 2007) on a randomly selected sample of 15% of live fidelity observations for each grade and intervention. The average interrater reliability for the reading and language components of the embedded and the standalone interventions exceeded 82% in all grades. In grades K–2, 72% to 91% of groups in the two interventions demonstrated at least 80% fidelity on the reading and language components. The median overall fidelity across interventions was 96% in kindergarten, 94% in grade 1, and 96% in grade 2.

The two fidelity ratings for each intervention group (i.e., fall reading and spring reading) were averaged to create an overall reading fidelity rating. Similarly, the two fidelity ratings for each intervention group (i.e., fall language and spring language) were averaged to create an overall language fidelity rating.

Program coverage. On average, interventionists covered 86% to 88% of the reading and language curricula in the embedded intervention and 77% to 79% of the language curricula in the stand-alone intervention across grades K–2. Interventionists covered 80% of the kindergarten lessons in *Sound Partners*. In the combined book for grades 1 and 2, interventionists covered, on average, 55% of the lessons in grade 1 and 62% of the lessons in grade 1. Covering half of the combined book in grade 1 is what would be expected. The relatively low coverage in grade 2 is likely due to the requirements for skill mastery in *Sound Partners*. Intervention groups across these grades required, on average, remediation on eight to 11 of 30 possible skill assessments. When remediation occurred, it was because an average of 49% to 59% of students in the intervention group had not mastered the skills taught. Therefore, half the group potentially benefited from needed remediation, and the other half received potentially unnecessary remediation.

Attendance. Each day, interventionists recorded students' attendance. In total, students could have attended approximately 134 days of intervention sessions. On average, students in the stand-alone intervention attended 92 to 95 days of intervention, and students in the embedded intervention attended 96 to 98 days of intervention across grades K–2.

In sum, both stand-alone and embedded interventions were implemented with high fidelity and good program coverage (except, perhaps, in *Sound Partners* in grade 2). Furthermore, attendance was high and attrition was low for both interventions. Because interventions did not differ, on average, in these variables indicative of the actual amount of intervention students received, these variables were not included in analyses of impact.

Measures

Table 2 reports the measures that were administered by grade, baseline, and outcome.

Florida Center for Reading Research Reading Assessment. The K–2 system of the FRA is a computer-adaptive screening assessment of reading and language, with normative information based on a large representative sample of Florida students (Foorman et al., 2015). Evidence of validity beyond content, concurrent, and predictive validity (Foorman et al., 2015) is provided by the fact that latent profiles of students found in the normative sample relate strongly to a reading comprehension factor composed of nationally normed tests (Foorman, Petscher, Stanley, & Truckenmiller, 2016). Finally, evidence of instructional utility for English-learner students was confirmed in a recent study (Foorman, Espinosa, Wood, & Wu, 2016).

The FRA can be administered up to three times a year. In all FRA tasks, students receive five items at grade level, and then the system adapts up or down based on performance to reach a precise estimate of a student's ability. The marginal reliability (Sireci, Thissen, & Wainer, 1991) for the FRA outcomes based on the nor-

Table 2. Measures Administered by Grade, Baseline, and Outcome

	Kindergarten		Grade 1		Grade 2	
Measure	Baseline	Outcome	Baseline	Outcome	Baseline	Outcome
FRA letter sounds	X					
FRA phonological awareness	X	X				
FRA word reading		X	X	X	X	X
FRA spelling					X	X
SESAT word reading		X				
FRA vocabulary pairs	X	X	X	X	X	X
FRA following directions	X	X	X	X	X	X
FRA sentence comprehension	X	X	X	X		X
SESAT sentence reading		X				
SAT-10 reading comprehension				X		X

Note.—FRA = Florida Center for Reading Research Reading Assessment; SESAT = Stanford Early Scholastic Achievement Test; SAT-10 = Stanford Achievement Test, Tenth Edition.

mative sample ranges from .85 to .96 across grades K–2. Tasks that span more than one grade have been vertically scaled across grades so that scores across grades K–2 are on the same scale. Students are given a developmental ability score on each task that has a mean of 500 and a standard deviation of 100.

Phonological awareness. This kindergarten task requires students to listen to a word that has been broken into parts and then blend them back together to reproduce the word. Sample-based, marginal reliability (Lord & Novick, 1968) was estimated at .75. Concurrent validity is provided by a correlation of .36 with the letter-word identification task of the Woodcock-Johnson III Test of Achievement (Woodcock, McGrew, & Mather, 2001).

Letter sounds. This kindergarten task requires students to provide the sound that a letter makes; the letter is presented on the computer in uppercase and lowercase. Sample-based, marginal reliability was estimated at .80. Concurrent validity is provided by a correlation of .52 with the phonemic awareness task of the Woodcock-Johnson III Test of Achievement (Woodcock et al., 2001).

Sentence comprehension. This kindergarten task requires students to select the one picture out of four presented on the computer that depicts the sentence given by the computer (e.g., click on the picture of "the bird flying towards the nest"). Concurrent validity is provided by a correlation of .48 with the sentence structure subtest from the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4; Semel, Wigg, & Secord, 2003). Sample-based, marginal reliability was estimated at .73.

Vocabulary pairs. This task was administered in grades K–2. Three words (or pictures, for kindergarten students) appear on the monitor and are pronounced by the computer. The student selects the two words that go together best (e.g., "dark," "night," "swim"). Concurrent validity is provided by correlations with the Peabody Picture Vocabulary Test, Fourth Edition (Dunn & Dunn, 2007) of .46 in kindergarten, .59 in grade 1, and .50 in grade 2. Sample-based, marginal reliability was estimated at .80 in kindergarten and grade 1 and at .70 in grade 2.

Following directions. This task was administered in grades K–2 and assesses listening comprehension, memory, and attention. Students listen and click and drag objects in response to the computer's directions (e.g., "Put the square in front of the chair and then put the circle behind the chair"). Concurrent validity is pro-

vided by correlations with the CELF-4 concepts and following directions subtest (Semel et al., 2003) of .58 in kindergarten, .58 in grade 1, and .68 in grade 2. Sample-based, marginal reliability was estimated at .83 in kindergarten, .80 in grade 1, and .73 in grade 2.

Word reading. This task was administered in grades 1 and 2. Words of varying difficulty (based on word frequency and item statistics) are presented on the monitor one at a time, and students read them aloud. Sample-based, marginal reliability was estimated at .91 in grade 1 and .92 in grade 2.

Spelling. This dictation task was administered in grade 2 only. The computer provides a word and uses it in a sentence. Students respond by using the computer keyboard to spell the word. Words reflected second-grade spelling patterns contained in state curriculum standards, the scope and sequence of spelling programs, and research on spelling development (see Arndt & Foorman, 2010, for more information). The bivariate correlation between spelling and word reading was .78, similar to what has been found in other studies (Arndt & Foorman, 2010). Sample-based, marginal reliability was estimated at .91.

Stanford Early Scholastic Achievement Test. The word reading and sentence reading subtests of the Stanford Early Scholastic Achievement Test (SESAT; Pearson Education, 2003) were used as an outcome for kindergarten students. The word reading subtest requires students to identify (*a*) the printed name for a picture of an object after the name has been pronounced, (*b*) the printed name for a picture of an object, (*c*) two printed words associated with a given picture, and (*d*) the printed word that has been pronounced. The Kuder-Richardson Formula 20 (KR20) reliability coefficient for spring administration is .85. The sentence reading subtest requires students to comprehend printed decodable sentences and sentences with decodable onset rime. The KR20 reliability coefficient for spring administration is .88.

Stanford Achievement Test, Tenth Edition, reading comprehension. The reading comprehension subtest of the SAT, Tenth Edition (SAT-10; Harcourt Brace, 2003) was used as an outcome for students in grades 1 and 2. At the Primary 1 test level, grade 1 students (*a*) identify the picture described by a two-sentence story, (*b*) read short selections and demonstrate explicit and implicit understanding by completing sentences in a modified cloze format, and (*c*) read short passages and answer multiple-choice questions tapping explicit and implicit information. The KR20 reliability coefficient for the Primary 1 is .91. In grade 2, students take the Primary 2 SAT-10. Students read literary, informational, and functional text passages and answer a total of 40 multiple-choice questions that assess initial understanding, interpretation, critical analysis, and awareness and usage of reading strategies. The KR20 reliability coefficient for the Primary 2 is .91. Validity was established with other standardized assessments of reading comprehension, providing evidence of content, criterion-related, and construct validity (Harcourt Assessment, 2004).

Data Analytic Plan

This section describes the multiple imputation procedures used to handle missing data, the analytic approach, and the procedure used to correct for multiple hypothesis testing. Prior to any data analysis, descriptive analyses were conducted to identify the presence of outliers and to verify that the data were normally distrib-

uted. Corrections for outliers were made during this data cleaning process, and all measures demonstrated normality. Outliers were identified using the median plus or minus two interquartile ranges, such that any value that exceeded this range was considered an outlier, and scores were changed to reflect the appropriate bound. Less than 5% of the total data points were identified as outliers across all grades (3.2% of the 12,732 data points in kindergarten, 3.7% of the 9,882 data points in grade 1, and 2.6% of the 13,090 data points in grade 2).

Multiple imputation for missing data. Multiple imputation for clustered data sets (Mistler, 2013) was used by grade, cohort, and intervention group to account for missing outcome data for students who moved between baseline and outcome and for whom no outcome data could be administered. The multiple imputation procedure was conducted using a multilevel multiple imputation macro in SAS (Mistler, 2013) that takes into account the nested structure of the data. In the imputation procedure, several variables including baseline, outcome, and student-level demographics (i.e., gender, free or reduced-price lunch status, English-learner status, and race/ethnicity) were used to inform the imputations. Overall, 1,000 imputed files per grade, cohort, and intervention group were created and aggregated for use in all analyses.

Multilevel analyses. For all research questions, a three-level hierarchical linear model (HLM), with students nested in small groups and then nested in schools was used to estimate treatment effects by grade using the MIXED procedure in SAS (Version 9.4). All analyses included student, small-group aggregated, and school aggregated FRA baseline measures as covariates. Cohort and region were also included as school-level covariates. All continuous predictors were grand mean centered. Prior to estimating any models, an unconditional model was estimated for each outcome to calculate the intraclass correlation for each level modeled in the estimated three-level HLM (see Table 3).

A top-down approach was used to answer the subresearch questions, in which the full subgroup model that included all covariates, the treatment indicator, and interactions (i.e., vectors of Baseline × Cohort, Baseline × Treatment, and Baseline × Cohort × Treatment interactions) was estimated first, and nonsignificant predictors (i.e., interactions and cohort) were then removed iteratively in subsequent models (West, Welch, & Galecki, 2007).[1]

The removal of nonsignificant predictors from the full subgroup model followed a systematic process, such that the three-way interactions (i.e., Baseline × Cohort × Treatment) were removed first, then two-way interactions (i.e., Baseline × Cohort and Baseline × Treatment), and finally, cohort. Baseline covariates at all levels were retained regardless of significance to increase the precision of the treatment effect. If the final subgroup model included a significant treatment interaction, the highest level interaction (i.e., the three- or two-way interaction) involving the treatment variable was explored further. A significant three-way interaction among treatment, cohort, and baseline was explored further by testing treatment differences within each cohort when baseline was 1 *SD* either above or below the mean. A significant two-way interaction between treatment and cohort was explored further by testing treatment differences within each cohort. Finally, a significant two-way interaction between treatment and baseline was explored further by testing treatment differences when baseline was 1 *SD* either above or below the mean.

Table 3. Percentage of Variance in Each Outcome between Students, Small Groups, and Schools by Grade

	SESAT/SAT-10			FRA					
Level	WR	SR	RC	PA	WR	SP	VP	FD	SC
Kindergarten:									
Student	72	75	–	91	68	–	93	90	87
Small group	8	12	–	3	8	–	5	5	1
School	20	13	–	6	24	–	2	5	12
Grade 1:									
Student	–	–	74	–	73	–	89	87	88
Small group	–	–	10	–	10	–	1	1	5
School	–	–	16	–	18	–	10	12	7
Grade 2:									
Student	–	–	81	–	83	82	91	87	94
Small group	–	–	6	–	5	10	1	0	0
School	–	–	13	–	12	8	8	13	6

Note.—SESAT = Stanford Early Scholastic Achievement Test; SAT-10 = Stanford Achievement Test, 10th Edition; FRA = Florida Center for Reading Research Reading Assessment; WR = word reading; SR = sentence reading; RC = reading comprehension; PA = phonological awareness; VP = vocabulary pairs; FD = following directions; SC = sentence comprehension.

The final subgroup model for each outcome by grade was then used as the base model when estimating treatment differences for English-learner and non-English-learner students. At a minimum, two variables were added to the base model: Student-level English-Learner Status and the cross-level English-Learner Status × Treatment interaction.

A Hedges's *g* effect size (Hedges, 1981) was calculated for all final models by dividing effect estimates by the unadjusted pooled within-group standard deviation. Hedges's *g* effect size differences of .25 or greater are highlighted as substantively important, following the WWC (2014) criteria.

Multiple hypothesis testing. The current study included multiple hypothesis tests by grade within two outcome domains: reading and language. The reading domain was represented by FRA measures of phonological awareness, word reading, and spelling and by SESAT word reading. The language domain was represented by FRA measures of following directions, vocabulary pairs, and sentence comprehension; SESAT sentence reading; and SAT-10 reading comprehension. The estimation of multiple hypothesis tests can increase the probability of falsely detecting a statistically significant treatment effect. Therefore, a correction must be applied to all significant treatment effects to reduce the false discovery rate. In the current study, the Benjamini-Hochberg linear step-up procedure (Benjamini & Hochberg, 1995) was used by research question, grade, and outcome domain to correct for multiple hypothesis testing.

Results

The first part of this section reports descriptive information by intervention and grade on baseline and outcome FRA percentile ranks. The second part of this section reports differences in outcome performance between the stand-alone and embedded interventions on measures of reading and language for at-risk students in

grades K–2. The final part of this section reports differences in outcome performance between the stand-alone and embedded interventions for each grade by cohort and baseline performance and, finally, by English-learner status.

Descriptive Analyses

Table 4 reports the baseline and outcome percentile ranks for the FRA measures by grade. On average, students in grades K–2 in both interventions started at or below the tenth percentile on the FRA reading tasks (i.e., phonological awareness in kindergarten, word reading in grades 1 and 2, and spelling in grade 2). By the end of the year, students had, on average, reached at least the twentieth percentile, with the exception of spelling in grade 2. The average difference between baseline and outcome for the FRA reading tasks ranged from 13 to 25 percentile points across grades.

In kindergarten, students started, on average, at or below the tenth percentile on FRA following directions and sentence comprehension and ended the year above the twenty-fifth percentile. The average difference between baseline and outcome for these FRA language tasks ranged from 20 to 25 percentile points. In grades 1 and 2, students in both interventions started, on average, between the tenth and fifteenth percentiles on FRA following directions and vocabulary pairs and ended the year between the eighteenth and thirtieth percentiles. The average difference between baseline and outcome for these FRA language tasks ranged from 6 to 15 percentile points.

The largest average percentile difference between baseline and outcome for any FRA task was in grade 1 for sentence comprehension. Students began just below

Table 4. Average Baseline and Outcome Percentile Rank and Difference by Intervention and Grade

	Stand-Alone			Embedded		
Outcome Measure	Baseline	Outcome	Difference	Baseline	Outcome	Difference
Kindergarten:						
FRA phonological awareness	1	21	20	1	26	25
FRA vocabulary pairs	25	34	9	24	33	9
FRA following directions	7	27	20	5	26	21
FRA sentence comprehension	10	35	25	9	32	23
Grade 1:						
FRA word reading	1	23	22	1	26	25
FRA vocabulary pairs	12	18	6	12	18	6
FRA following directions	10	19	9	11	21	10
FRA sentence comprehension	29	64	35	27	66	39
Grade 2:						
FRA word reading	5	24	19	9	26	17
FRA spelling	3	22	19	4	17	13
FRA vocabulary pairs	12	22	10	10	18	8
FRA following directions	15	30	15	13	26	13

Note.—Percentile ranks are based on winter norms. The sentence comprehension task is a kindergarten normed assessment, which means that the percentile ranks for all grades are reflective of ability on a kindergarten scale. FRA = Florida Center for Reading Research Reading Assessment.

the thirtieth percentile in both the stand-alone and embedded interventions and ended the year above the sixtieth percentile. This reflects an average difference between baseline and outcome from 35 to 39 percentile points across interventions. However, it is important to note that the norms for this task are based on kindergarten students and thus are reflective of ability on a kindergarten scale.

Primary Impacts of the Two Interventions

Prior to estimating any treatment effects, we investigated baseline equivalence for the analytic sample by grade, cohort, and full sample (see Tables 5 and 6). From Tables 5 and 6, it is readily apparent that Cohort 2 schools were lower performing than Cohort 1 schools; therefore, cohort needed to be retained in analyses. The likely reason for cohort differences was that changes in the state test after Cohort 1 triggered changes in the school grading system that led to an increase in schools with low school grades. The majority of baseline differences between the two interventions were determined to be nonsignificant across baseline measures, cohort, and grade. However, a significant difference between the two interventions was observed on FRA following directions in kindergarten for Cohort 1, $t(465) = 2.64$, $p = .01$; FRA word reading in grade 1 for Cohort 1, $t(503) = -3.02$, $p = .003$; FRA vocabulary pairs in grade 1 for Cohort 1, $t(503) = -2.16$, $p = .03$; FRA word reading in grade 2 for Cohort 1, $t(621) = -2.82$, $p = .005$; and FRA vocabulary pairs in grade 2 for Cohort 2, $t(661) = 3.03$, $p = .003$. For the full sample, the only significant differ-

Table 5. Baseline Differences between the Stand-Alone and Embedded Interventions for the Analytic Sample on FRA Reading Outcomes by Grade, Cohort, and the Full Sample

	Letter Sounds			Phonological Awareness or Word Reading[a]			Spelling		
Sample	Stand-Alone *M* (*SD*)	Embedded *M* (*SD*)	ES	Stand-Alone *M* (*SD*)	Embedded *M* (*SD*)	ES	Stand-Alone *M* (*SD*)	Embedded *M* (*SD*)	ES
Kindergarten:									
Cohort 1	335 (91)	347 (106)	−.12	312 (113)	292 (122)	.17	–	–	–
Cohort 2	272 (99)	267 (102)	−.94	246 (84)	253 (86)	−.08	–	–	–
Full	304 (100)	303 (111)	.01	279 (105)	271 (103)	.08	–	–	–
Grade 1:									
Cohort 1	–	–	–	257 (186)	342 (205)	−.43**	–	–	–
Cohort 2	–	–	–	231 (139)	253 (148)	−.15	–	–	–
Full	–	–	–	243 (165)	293 (179)	−.29**	–	–	–
Grade 2:									
Cohort 1	–	–	–	478 (72)	521 (94)	−.52**	351 (106)	357 (107)	−.06
Cohort 2	–	–	–	457 (102)	453 (103)	.04	318 (98)	321 (99)	.02
Full	–	–	–	469 (90)	484 (105)	−.15	338 (103)	338 (104)	.00

Note.—Baseline means are adjusted using geographic region as a covariate. *SD* = unadjusted standard deviation; ES = effect size estimated using Hedges's *g*.

[a] Phonological awareness was administered to kindergarten students at baseline, and word reading was administered to grade 1 and grade 2 students at baseline.

** $p < .01$.

Table 6. Baseline Differences between the Stand-Alone and Embedded Interventions for the Analytic Sample on FRA Language Outcomes by Grade, Cohort, and the Full Sample

	Vocabulary Pairs			Following Directions			Sentence Comprehension		
Sample	Stand-Alone *M* (*SD*)	Embedded *M* (*SD*)	ES	Stand-Alone *M* (*SD*)	Embedded *M* (*SD*)	ES	Stand-Alone *M* (*SD*)	Embedded *M* (*SD*)	ES
Kindergarten:									
Cohort 1	364 (57)	366 (60)	−.03	286 (127)	243 (156)	.30**	409 (90)	397 (92)	.13
Cohort 2	339 (65)	335 (65)	.06	237 (154)	233 (149)	.03	398 (85)	399 (82)	−.01
Full	351 (63)	349 (64)	.03	261 (144)	237 (152)	.16	403 (87)	399 (86)	.05
Grade 1:									
Cohort 1	412 (62)	425 (65)	−.20*	408 (115)	431 (119)	−.20	462 (129)	451 (125)	−.09
Cohort 2	409 (71)	399 (74)	.14	398 (116)	386 (124)	.09	465 (70)	464 (77)	.01
Full	411 (67)	410 (71)	.02	402 (115)	407 (124)	−.04	463 (104)	458 (100)	.05
Grade 2:									
Cohort 1	497 (82)	503 (87)	−.07	535 (113)	433 (103)	.02	–	–	–
Cohort 2	500 (67)	475 (72)	.36**	490 (119)	476 (127)	.11	–	–	–
Full	499 (75)	488 (81)	.15	514 (118)	502 (120)	.10	–	–	–

Note.—Baseline means are adjusted using geographic region as a covariate. *SD* = unadjusted standard deviation; ES = effect size estimated using Hedges's *g*.

* $p < .05$.

** $p < .01$.

ence between the two interventions was observed in grade 1 on FRA word reading, $t(1095) = -2.61$, $p = .01$. All baseline measures were used as covariates in all models.

Table 7 reports the adjusted means and unadjusted standard deviations for reading and language outcomes by intervention and grade, as well as the *p*-values and Hedges's *g* effect size differences comparing the stand-alone and embedded interventions. The full-sample HLM models revealed nonsignificant treatment effects between the stand-alone and embedded interventions for all reading and language outcomes in all grades, with the exception of FRA spelling in grade 2. In grade 2, the stand-alone intervention significantly improved FRA spelling outcome scores relative to the embedded intervention by 17 ability score points (see Table 7). This statistically significant 17-point difference reflects an effect size of $g = .18$.

Differences Between Interventions Based on Cohort and Baseline Performance

The majority of outcomes within each grade did not include a significant interaction involving the treatment indicator in the final subgroup HLM model, including FRA phonological awareness in kindergarten, FRA word reading in kindergarten and grade 2, FRA vocabulary pairs in all grades, FRA following directions in kindergarten and grade 2, FRA sentence comprehension in kindergarten and grade 1, SESAT sentence reading in kindergarten, and SAT-10 reading comprehension in grades 1 and 2. Table 8 reports the adjusted means and unadjusted standard deviations by intervention and grade, and the *p*-values and Hedges's *g* effect size differences between the stand-alone and embedded interventions for reading and language outcomes, demonstrating a significant Treatment × Baseline, Treatment ×

Table 7. Primary Impacts of the Stand-Alone and Embedded Interventions by Grade and Outcome

Outcome	Adjusted Mean (*SD*)		Difference (*SE*)	*p*	ES
	Stand-Alone	Embedded			
Kindergarten:					
FRA phonological awareness	434 (147)	452 (134)	−18 (13)	.18	−.13
FRA word reading	332 (134)	337 (149)	−5 (17)	.79	−.04
SESAT word reading	433 (38)	426 (35)	7 (5)	.18	.19
FRA vocabulary pairs	369 (77)	372 (76)	−3 (5)	.56	−.04
FRA following directions	358 (115)	363 (105)	−5 (7)	.56	−.05
FRA sentence comprehension	472 (81)	476 (77)	−4 (6)	.58	−.05
SESAT sentence reading	459 (50)	460 (46)	−1 (6)	.80	−.02
Grade 1:					
FRA word reading	448 (105)	436 (123)	12 (13)	.37	.10
FRA vocabulary pairs	435 (79)	428 (86)	7 (7)	.35	.08
FRA following directions	442 (109)	440 (117)	2 (9)	.80	.02
FRA sentence comprehension	542 (87)	542 (87)	0 (6)	.96	.00
SAT-10 reading comprehension	519 (39)	514 (42)	5 (5)	.28	.12
Grade 2:					
FRA word reading	546 (82)	541 (100)	5 (8)	.52	.05
FRA spelling	434 (88)	417 (98)	17 (6)	.009[a]	.18
FRA vocabulary pairs	526 (80)	519 (81)	8 (6)	.23	.09
FRA following directions	556 (122)	548 (125)	8 (9)	.31	.05
FRA sentence comprehension	601 (89)	589 (88)	12 (7)	.06	.10
SAT-10 reading comprehension	565 (31)	565 (32)	0 (3)	.88	.00

Note.—SD = unadjusted standard deviation; SE = standard error; ES = effect size estimated using Hedges's *g*; FRA = Florida Center for Reading Research Reading Assessment; SESAT = Stanford Early Scholastic Achievement Test; SAT-10 = Stanford Achievement Test, Tenth Edition.

[a] This *p*-value is significant after applying the Benjamini-Hochberg correction procedure (Benjamini & Hochberg, 1995), where the identified *p*-value cutoff for reading outcomes is $p \leq .025$.

Cohort, or Treatment × Baseline × Cohort interaction in the final subgroup HLM model.

In kindergarten, a significant FRA sentence comprehension Baseline × Cohort × Treatment interaction was observed on the SESAT word reading outcome, $\gamma 102 = -.098$, $t(1{,}038) = -2.16$, $p = .03$. This interaction was further explored by investigating differences between the interventions for each cohort at ±1 *SD* from the FRA sentence comprehension baseline grand mean. Results from this investigation did not show any significant comparisons; however, they highlighted two substantively important differences (effect size greater than .25) between the two interventions in favor of the stand-alone intervention. Specifically, the stand-alone intervention resulted in substantively higher SESAT word reading outcome scores in Cohort 1 for students with FRA sentence comprehension baseline scores 1 *SD* above the grand mean ($g = .37$) and for students in Cohort 2 with FRA sentence comprehension baseline scores 1 *SD* below the grand mean ($g = .28$) compared with similar students in the embedded intervention (see Table 8). As shown in Table 6, Cohort 2 students in the stand-alone intervention were lower at baseline in sentence comprehension than Cohort 1 students. It is noteworthy that it was stand-alone students 1 *SD* below the grand mean in sentence comprehension who were boosted more in word reading skills relative to similarly low-baseline students in the embedded intervention (see Table 8).

Table 8. Differences in Adjusted Means between the Stand-Alone and Embedded Interventions for Outcomes Demonstrating Significant Interactions by Grade

		Adjusted Mean (*SD*)				
Outcome	Subgroup	Stand-Alone	Embedded	Difference (*SE*)	*p*	ES
Kindergarten:						
SESAT WR	Cohort 1 high FRA SC baseline	435 (39)	421 (33)	14 (8)	.09	.37
SESAT WR	Cohort 1 low FRA SC baseline	426 (39)	418 (33)	8 (8)	.34	.22
SESAT WR	Cohort 2 high FRA SC baseline	435 (36)	437 (35)	−2 (7)	.76	−.06
SESAT WR	Cohort 2 low FRA SC baseline	435 (36)	425 (35)	10 (7)	.22	.28
Grade 1:						
FRA WR	Cohort 1	461 (63)	483 (89)	−22 (17)	.19	−.29
FRA WR	Cohort 2	433 (131)	411 (129)	22 (15)	.13	.17
FRA FD	Low FRA WR baseline	433 (109)	445 (117)	−12 (10)	.27	−.11
FRA FD	High FRA WR baseline	453 (109)	435 (117)	−18 (11)	.11	.16
FRA FD	Low FRA FD baseline	402 (109)	385 (117)	−17 (11)	.09	.15
FRA FD	High FRA FD baseline	484 (109)	495 (117)	−11 (11)	.26	−.10
Grade 2:						
FRA SP	Low FRA SP baseline	404 (89)	379 (98)	25 (8)	.001[a]	.27
FRA SP	High FRA SP baseline	462 (89)	453 (98)	9 (8)	.29	.10
FRA SC	Cohort 1 low FRA VP baseline	597 (99)	559 (101)	38 (12)	.001[a]	.38
FRA SC	Cohort 1 high FRA VP baseline	607 (99)	589 (101)	18 (11)	.12	.18
FRA SC	Cohort 2 low FRA VP baseline	585 (76)	594 (75)	−9 (12)	.71	−.12
FRA SC	Cohort 2 high FRA VP baseline	616 (76)	603 (75)	13 (12)	.28	.18

Note.—Significant interactions involving the treatment indicator (i.e., Baseline × Treatment, Cohort × Treatment, or Baseline × Cohort × Treatment) were probed further and were included in this table. Significant Baseline × Cohort × Treatment interactions were probed at ±1 *SD* and are described as high or low in the subgroup column. If the final subgroup model for an outcome did not include any significant interactions involving the treatment indicator, it was excluded from this table. *SD* = unadjusted standard deviation; *SE* = standard error; ES = effect size estimated using Hedges's *g*; SESAT = Stanford Early Scholastic Achievement Test; WR = word reading; FRA = Florida Center for Reading Research Reading Assessment; SC = sentence comprehension; FD = following directions; SP = spelling; VP = vocabulary pairs.

[a] This *p*-value is significant after applying the Benjamini-Hochberg correction procedure (Benjamini & Hochberg, 1995), where the identified *p*-value cutoff for FRA SP is $p \leq .0025$ and for FRA SC is $p \leq .00125$.

In grade 1, a significant Cohort × Treatment interaction was observed on the FRA word reading outcome, $\gamma 005 = .057$, $t(1{,}080) = 2.46$, $p = .01$. This interaction was explored further by investigating differences between the interventions for each cohort. Results from this investigation did not show any significant comparisons; however, a substantively important 22-point difference on the FRA word reading outcome in favor of the embedded intervention in Cohort 1 ($g = .29$) was observed. In addition, a significant FRA word reading Baseline × Treatment interaction, $\gamma 102 = .008$, $t(1{,}081) = 2.36$, $p = .02$, and a significant FRA following directions Baseline × Treatment interaction, $\gamma 102 = -.001$, $t(1{,}081) = -2.41$, $p = .02$, were observed for the FRA following directions outcome. Further investigation of these interactions revealed no significant or substantively important differences between the two interventions at ±1 *SD* from the grand mean of FRA word reading baseline or FRA following directions baseline on the FRA following directions outcome.

In grade 2, a significant FRA spelling Baseline × Treatment interaction was observed on the FRA spelling outcome, $\gamma 102 = .0008$, $t(1{,}272) = 2.00$, $p = .045$. Further investigation of this interaction showed a significant improvement in FRA spelling scores for grade 2 students in the stand-alone intervention, with FRA spelling baseline scores 1 *SD* below the mean compared with similar students in

the embedded intervention (see Table 8). The average estimated FRA spelling ability score for students with low FRA spelling baseline scores in the stand-alone intervention was 404, compared with 379 in the embedded intervention. This statistically significant 25-point difference reflects an effect size of $g = .27$. This finding remained significant after applying the Benjamini-Hochberg correction.

A significant FRA vocabulary pairs Baseline × Cohort × Treatment interaction was also observed in grade 2 on FRA sentence comprehension outcome, $\gamma 103 = .03$, $t(1,268) = 2.15$, $p = .03$. Further investigation of this interaction resulted in a significant 38-point difference on FRA sentence comprehension in favor of the stand-alone intervention for grade 2 students in Cohort 1 with FRA vocabulary pairs baseline scores 1 *SD* below the grand mean ($g = .38$; see Table 8). This finding remained significant after applying the Benjamini-Hochberg correction.

Differences Between and Within Interventions for English-Learner and Non-English-Learner Students

No significant English-Learner Status × Treatment interaction was observed on any reading (i.e., FRA word reading and FRA spelling) or language (i.e., FRA vocabulary pairs, FRA following directions, FRA sentence comprehension, and SAT-10 reading comprehension) outcomes in grades 1 and 2 or language (i.e., FRA vocabulary pairs, FRA following directions, FRA sentence comprehension, and SESAT sentence reading) outcomes in kindergarten. However, a significant English-Learner Status × Treatment interaction was observed for all reading outcomes in kindergarten. Each significant interaction was explored further. Table 9 reports the adjusted means and unadjusted standard deviations for the kindergarten reading outcomes, as well as the *p*-values and Hedges's *g* effect sizes comparing differences in outcomes between interventions for English-learner and non-English-learner students. Table 10 reports the adjusted means and unadjusted standard deviations within each intervention for English-learner and non-English-learner students.

Three substantively important effects on reading outcomes were found in kindergarten. First, kindergarten English-learner students in the embedded intervention

Table 9. Differences in Reading Outcomes between Stand-Alone and Embedded Interventions for Kindergarten English-Learner and Non-English-Learner Students

		Sample Size		Adjusted Mean (*SD*)				
Outcome	Study Sample	Stand-Alone	Embedded	Stand-Alone	Embedded	Difference (*SD*)	*p*	ES
FRA PA	Non-EL	343	297	435 (147)	439 (131)	−4 (15)	.79	−.03
FRA PA	EL	169	213	425 (146)	470 (138)	−45 (18)	.01[a]	−.32
FRA WR	Non-EL	343	297	327 (138)	354 (151)	−27 (17)	.11	−.19
FRA WR	EL	169	213	333 (129)	328 (146)	5 (14)	.69	.04
SESAT WR	Non-EL	343	297	433 (37)	422 (34)	11 (5)	.04[a]	.31
SESAT WR	EL	169	213	429 (41)	431 (36)	−2 (6)	.76	−.05

Note.—*SD* = unadjusted standard deviation; ; ES = effect size estimated using Hedges's *g*; FRA = Florida Center for Reading Research Reading Assessment; PA = phonological awareness; EL = English-learner students; WR = word reading; SESAT = Stanford Early Scholastic Achievement Test.

[a] This *p*-value is not significant after applying the Benjamini-Hochberg correction procedure (Benjamini & Hochberg, 1995), where the identified *p*-value cutoff is $p \leq .004$.

Table 10. Differences in Reading Outcomes between Kindergarten English-Learner and Non-English-Learner Students within the Stand-Alone or Embedded Intervention

		Sample Size		Adjusted Mean (*SD*)				
Outcome	Intervention	Non-EL Students	EL Students	Non-EL Students	EL Students	Difference (*SE*)	*p*	ES
FRA PA	Embedded	297	213	439 (131)	470 (138)	−31 (14)	.02[a]	−.23
FRA PA	Stand-alone	343	169	435 (147)	425 (146)	10 (15)	.50	.07
FRA WR	Embedded	297	213	328 (146)	354 (151)	−26 (12)	.03[a]	−.18
FRA WR	Stand-alone	342	169	333 (129)	327 (138)	6 (13)	.61	.04
SESAT WR	Embedded	297	213	422 (34)	431 (36)	−9 (3)	.006[a]	−.27
SESAT WR	Stand-alone	343	169	433 (37)	429 (41)	4 (4)	.28	.10

Note.—*SD* = unadjusted standard deviation; EL = English-learner; *SE* = standard error; ES = effect size estimated using Hedges's *g*; FRA = Florida Center for Reading Research Reading Assessment; PA = phonological awareness; WR = word reading; SESAT = Stanford Early Scholastic Achievement Test.

[a] This *p*-value is not significant after applying the Benjamini-Hochberg correction procedure (Benjamini & Hochberg, 1995), where the identified *p*-value cutoff is $p \leq .004$.

performed substantively better than English-learner students in the stand-alone intervention on FRA phonological awareness by 45 points ($g = .32$; see Table 9). Second, a substantively important effect on SESAT word reading was observed for kindergarten non-English-learner students in favor of the stand-alone intervention relative to the embedded intervention (see Table 9). Specifically, non-English-learner students in the stand-alone intervention scored approximately 11 points higher on SESAT word reading than non-English-learner students in the embedded intervention ($g = .31$). Third, kindergarten English-learner students in the embedded intervention performed substantively better on SESAT word reading compared with non-English-learner students in the same intervention ($g = .27$; see Table 10).

Discussion

This study compared a small-group Tier 2 intervention embedded within the core reading/language arts program with an evidence-based, small-group Tier 2 stand-alone intervention outside of the core reading/language arts program in 55 very low-performing schools in Florida. REL Southeast staff compiled the materials for both interventions and trained interventionists who were either school staff or locally hired personnel. Thus, this study contrasted two plausible Tier 2 interventions rather than contrasting an experimental intervention against a business-as-usual control, which in these schools meant no intervention in grades K–2. The interventions in this study were implemented with fidelity and good coverage for 45 minutes daily throughout the school year. Student attendance was high and attrition was low.

Students in grades K–2, on average, showed improvement in reading and language skills in both interventions, starting the school year below the tenth percentile on the FRA phonological awareness and word reading measures and ending the year above the twentieth percentile. Kindergarten students in both intervention groups improved by 24 percentile points on the FRA sentence comprehension—the kind of listening comprehension task used to evaluate English-language profi-

ciency for the English-learner students in this study (Educational Testing Service, 2005). In addition, kindergarten students scored between the twentieth and twenty-sixth percentiles on the SESAT word reading and sentence reading outcomes at the end of the year across both interventions and cohorts.

The average gains on FRA reading-related tasks of 13 to 25 percentile points across grades compare well with the gains of 13 to 17 percentile points on word reading and pseudoword reading reported by Gersten and colleagues (Gersten, Jayanthi, et al., 2017; Gersten, Newman-Gonchar, et al., 2017) in their systematic review of Tier 2 studies in grades 1 to 3. In addition, these results alleviate concerns in this study that observed gains are caused by regression to the mean or expected growth due to classroom instruction. In fact, the reading-related gains demonstrated by the stand-alone and embedded interventions are noteworthy, given that the group size in the current study was four to five students compared with the average group sizes of Gersten and colleagues (Gersten, Jayanthi, et al., 2017; Gersten, Newman-Gonchar, et al., 2017), with 3.4 in grade 1 and 2.8 in grade 2. With respect to reading comprehension outcomes in the current study, word reading skills for students in grades 1 and 2 were not sufficiently developed, on average, to achieve reading comprehension outcomes above the fifteenth percentile.

Relative Impacts of the Stand-Alone and Embedded Interventions

The overarching research question in the current study asked whether there were differential impacts due to intervention. Reading and language outcomes were comparable in stand-alone and embedded intervention schools, except that stand-alone schools had significantly improved FRA spelling outcomes in grade 2 (effect size of .18). Grade 2 was the only grade with a spelling outcome, and the reading component of the stand-alone intervention (i.e., *Sound Partners*), unlike the embedded intervention, taught spelling by having students write the words they learned to read. By learning to encode (i.e., spell) as well as decode the words taught, *Sound Partners* is similar to other early reading interventions with significant impacts on reading outcomes (e.g., Foorman, Beyler, et al., 2016; Weiser & Mathes, 2011). However, in this study, the primary impacts for the stand-alone intervention relative to the embedded intervention were only on spelling and did not generalize to reading. This finding differs from the significantly positive effects found with *Sound Partners* by Vadasy and Sanders (2012) in grade 1 on word and pseudoword reading (effect size of .51) and passage reading fluency (effect size of .69) but is similar to other studies of *Sound Partners* with potentially positive results in grade 1 (Jenkins, Peyton, Sanders, & Vadasy, 2004) and in grades 2 and 3 (Vadasy et al., 2006; Vadasy, Sanders, & Tudor, 2007). In these studies, however, *Sound Partners* was delivered one on one—a big difference from the group size of four to five students in the current study.

Differential Impacts of Stand-Alone and Embedded Interventions by Baseline and Cohort

Within the overarching research question regarding relative impacts of the two interventions, there was a subquestion about differential impacts due to cohort or baseline. Cohort 2 schools were lower performing at baseline, probably due to the

change in Florida's school grading system with the advent of a more rigorous state test. Consequently, there were some inconsistencies in the pattern of relative effects of the two interventions by cohort and baseline scores across all grades. Several interactions favored the stand-alone intervention. First, across cohorts, the stand-alone intervention resulted in significantly improved spelling outcomes relative to the embedded intervention among students with low baseline spelling scores (effect size of .27). Second, the stand-alone intervention also had one significant effect relative to the embedded intervention on sentence comprehension in grade 2 in Cohort 1 for students with low baseline vocabulary scores (effect size of .38) and two substantively important effects on the SESAT word reading outcome in kindergarten—one in Cohort 1 for students with high baseline sentence comprehension scores (effect size of .37) and one in Cohort 2 for students with low baseline sentence comprehension scores (effect size of .28). Inconsistent with these results is the finding that the embedded intervention resulted in a substantially improved FRA word reading outcome relative to the stand-alone intervention among students in Cohort 1 schools in grade 1 (effect size of .29). This effect was not significant in the lower performing Cohort 2 schools. The lack of a consistent pattern of intervention effects across cohorts (except for spelling) implies that, on average, improvement was relatively comparable among students in schools in both intervention groups.

Differential Impacts of Stand-Alone and Embedded Interventions by English-Learner Status

There were no differences in reading and language outcomes in grades 1 and 2 or in language outcomes in kindergarten between English-learner students and non-English-learner students in schools in the same intervention group. Within non-English learners in kindergarten, students in the stand-alone intervention performed substantially better on SESAT word reading than their peers in the embedded intervention. In contrast, non-English learners in kindergarten did not improve as much as their English-learner peers on SESAT word reading within the embedded intervention. In addition, the FRA phonological awareness outcome among kindergarten English-learner students was substantially higher in embedded intervention schools than in stand-alone intervention schools.

Both interventions included instruction in phonological awareness, but the addition of comprehension activities in the embedded intervention may have helped scaffold English-learner students' ability to segment sounds in speech and to decode sight words. This finding is consistent with studies showing an advantage in phonological awareness tasks for bilingual students (e.g., Bialystok, Majumder, & Martin, 2003). These results also underscore the value of emphasizing comprehension when building on English-learner students' sensitivity to sounds in speech to connect to the sound-spelling patterns fundamental to reading. Finally, the fact that the non-English-learner students in stand-alone intervention schools scored higher on the SESAT word reading outcome than did their peers in embedded intervention schools suggests that the decontextualized nature of alphabetic instruction in *Sound Partners* was sufficient to build their word reading skills.

Conclusion and Implications for Future Research

The relatively few, small impacts (with effect sizes ranging from .18 to .38) raise the question of the significance of this study. The significance is that two different approaches to selection of instructional materials for small-group, Tier 2 intervention can be effective in improving reading and language skills in at-risk students in grades K–2 in low-performing schools. Both interventions were more similar than different in core components, and it was primarily in the differences in core components of spelling, inferential language, and comprehension where differential effects were found.

The study is also significant because the Tier 2 materials embedded within the core reading/language arts program, and thus available at no extra cost, can be as effective as stand-alone, evidence-based, Tier 2 materials purchased separately from the core reading/language arts program. This finding is consistent with the results of studies conducted by Fien and colleagues (2015; Smith et al., 2016) on the explicit, systematic ECRI in which Tier 2 instruction is aligned with the scope and sequence of Tier 1 instruction.

Alignment of Tier 2 instruction with Tier 1 instruction was important to the classroom teachers in the current study because Florida teachers are evaluated on the performance of all students in their classrooms, including those pulled out for reading intervention. However, an effective implementation of the Tier 2 reading and language components of HMH *Journeys* requires that someone at the school compile the pieces of Strategic Intervention—information from the tab in the teacher's edition, the assessment book, and the activity cards—and order the supplementary vocabulary program Curious about Words, as was done in the current study by REL Southeast staff. In addition to compiling these embedded Tier 2 materials into an implementation manual for interventionists, someone at the school needs to be responsible for training interventionists and providing ongoing support to ensure fidelity.

Finally, future experiments could modify the stand-alone intervention in ways that might make it easier to implement. First, it was challenging for interventionists to decide how to remediate students in *Sound Partners* on different skills and what to do with students who did not need remediation. It is not surprising that *Sound Partners* was delivered one on one in studies with positively significant effects (Vadasy & Sanders, 2012) and potentially positive effects (Jenkins et al., 2004; Vadasy et al., 2007). If *Sound Partners* were to be delivered in small groups in a future version of the stand-alone intervention, perhaps it could be delivered without remediation and contrasted with the current version to see whether student reading outcomes differed. Second, interventionists in the current study had to remember which day to teach vocabulary and which day to teach inferential language during the week. This was challenging because of occasional changes in schedules due to events such as school plays that required interventionists to remember which language piece had to be rescheduled for each of their intervention groups. An integrated version of the language component in the stand-alone intervention where vocabulary and inferential language are taught each day could be contrasted with the current version to see whether student language outcomes differed.

Notes

This study was supported by the Institute of Education Sciences, U.S. Department of Education, under contract ED-IES-12-C-0011 to Florida State University for the Regional Educational Laboratory Southeast. The opinions expressed are those of the authors and do not represent views of the institute, the U.S. Department of Education, or Florida State University. Barbara R. Foorman is the Francis Eppes Professor of Education and director emeritus of the Florida Center for Reading Research at Florida State University; Sarah Herrera is an associate in research at the Florida Center for Reading Research at Florida State University; Jennifer Dombek is an associate in research at the Florida Center for Reading Research at Florida State University. Correspondence may be sent to Barbara R. Foorman, Florida Center for Reading Research, Florida State University, 2010 Levy Avenue, Suite 100, Tallahassee, FL 32310; bfoorman@fcrr.org.

1. Results from the full subgroup model are available on request from the authors.

References

Arndt, E., & Foorman, B. (2010). Second graders as spellers: What types of errors are they making? *Assessment for Effective Intervention*, **36**(1), 57–67. doi:10.1177/1534508410380135

Baker, S., Fien, H., & Baker, D. (2010). Robust reading instruction in the early grades: Conceptual and practical issues in the integration and evaluation of Tier 1 and Tier 2 instructional supports. *Focus on Exceptional Children,* **42**(9), 1–20.

Baker, S., Lesaux, N., Jayanthi, M., Dimino, J., Proctor, C. P., Morris, J., . . . Newman-Gonchar, R. (2014). *Teaching academic content and literacy to English learners in elementary and middle school* (NCEE 2014-4012). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/Docs/practiceguide/english_learners_pg_040114.pdf

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). *Evaluation of response to intervention practices for elementary school reading* (NCEE No. 2016-4000). Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/pubs/20164000/pdf/20164000.pdf

Beck, I., McKeown, M., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**(1), 289–300. Retrieved from http://www.jstor.org/stable/2346101

Bialystok, E., Majumder, S., & Martin, M. M. (2003). Developing phonological awareness: Is there a bilingual advantage? *Applied Psycholinguistics*, **24**, 27–44. doi:10.1017.S014271640300002X

Catts, H., Nielsen, D. C., Bridges, M. S., & Liu, Y. (2014). Early identification of reading comprehension difficulties. *Journal of Learning Disabilities*, **49**(5), 451–465. doi:10.1177/0022219414556121

Dunn, L., & Dunn, D. (2007). *Peabody Picture Vocabulary Test–4*. San Antonio: Pearson.

Educational Testing Service. (2005). *Comprehensive English Language Learning Assessment (CELLA): Technical summary report*. Princeton, NJ: Authors. Retrieved from http://www.accountabilityworks.org/photos/CELLA_Technical_Summary_Report.pdf

Fien, H., Smith, J., Smolkowski, K., Baker, S., Nelson-Walker, N., & Chaparro, E. (2015). An examination of the efficacy of a multitiered intervention on early reading outcomes for first grade students at risk for reading difficulties. *Journal of Learning Disabilities,* **48**, 602–621.

Fletcher, J., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, *3*, 30–37. doi:10.1111/j.1750-8606.2008.00072.x

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., . . . Wissell, S. (2016). *Foundational skills to support reading for understanding in kindergartenthrough 3rd grade*. Washington, DC: National Center for Education Evaluation and Regional Assistance,

Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/Docs/practiceguide/wwc_foundationalreading_070516.pdf

Foorman, B., Espinosa, A., Wood, C., & Wu, Y. (2016). *Using computer-adaptive literacy assessments to monitor the progress of English language learner students* (REL 2016-149). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_2016149.pdf

Foorman, B., Herrera, S., Dombek, J., Schatschneider, C., & Petscher, Y. (2017). *The relative effectiveness of two approaches to early literacy intervention in grades K–2* (REL 2017-251). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from https://ies.ed.gov/ncee/edlabs/regions/southeast/pdf/REL_2017251.pdf

Foorman, B., Petscher, Y., & Schatschneider, C. (2015). *Florida Center for Reading Research (FCRR) Reading Assessment: Technical manuals*. Tallahassee: Florida Center for Reading Research. Retrieved from http://www.fcrr.org/for-researchers/fra.asp

Foorman, B., Petscher, Y., Stanley, C., & Truckenmiller, A. (2016). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness*, **10**(3), 619–645. doi:10.1080/19345747.2016.1237597

Foorman, B. R., & Connor, C. (2011). Primary reading. In M. Kamil, P. D. Pearson, & E. Moje (Eds.), *Handbook on reading research* (Vol. **4**, pp. 136–156). New York: Taylor & Francis.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., . . . Tilly, W. D. (2009). *Assisting students struggling with reading: Response to intervention and multi-tier intervention for reading in the primary grades: A practice guide* (NCEE No. 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/rti_reading_pg_021809.pdf

Gersten, R., Jayanthi, M., & Dimino, J. (2017). Too much, too soon? A commentary on what the national response-to-intervention evaluation left unanswered and what reading intervention research tells us. *Exceptional Children*, **83**, 244–254.

Gersten, R., Newman-Gonchar, R., Haymond, K., & Dimino, J. (2017). *What is the evidence base for response to intervention in reading in grades 1–3?* (REL 2016-129). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID = 4558

Harcourt Assessment. (2004). *Stanford technical data report*. San Antonio: Author.

Harcourt Brace. (2003). *Stanford Achievement Test* (10th ed.). San Antonio: Author.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, **1**(1), 77–89.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107–128. doi:10.3102/10769986006002107

Jenkins, J., Peyton, J., Sanders, E., & Vadasy, P. (2004). Effects of reading decodable text in supplemental first-grade tutoring. *Scientific Studies of Reading*, **8**(1), 53–85. doi:10.1207/s1532799xssr0801_4

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

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mistler, S. A. (2013). *A SAS macro for applying multiple imputation to multilevel data* (Paper 438-2013). Retrieved from https://support.sas.com/resources/papers/proceedings13/438-2013.pdf

Montgomery, J. (2007). *Bridge of vocabulary*. New York: Pearson.

National Institute of Child Health and Human Development. (2000). *National reading panel—Teaching children to read: Reports of the subgroups* (NIH Pub. No. 00-4754). Washington,

DC: U.S. Department of Health and Human Services. Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/Documents/report.pdf

Pearson Education. (2003). *Stanford Early Scholastic Achievement Test (SESAT)*. New York: Author.

Phillips, B. (2014). Promotion of syntactical development and oral comprehension: Development and initial evaluation of a small-group intervention. *Child Language Teaching and Therapy*, **30**(1), 63–77. doi:10.1177/0265659013487742

Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., & Seidenberg, M. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, **2**(2), 31–74.

Semel, E., Wigg, E., & Secord, W. (2003). *The Clinical Evaluation of Language Fundamentals* (4th ed.): *Examiner's manual*. San Antonio: Pearson.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, **28**, 237–247. doi:10.1111/j.1745-3984.1991.tb00356.x

Smith, J., Smolkowski, K., Baker, S., Fien, H., & Kosty, D. (2016). Examining the efficacy of a multitiered intervention for at-risk readers in grade 1. *Elementary School Journal*, **116**(4), 549–573. doi:10.1086/686249

Vadasy, P., & Sanders, E. (2012). Two-year follow-up of a kindergarten phonics intervention for English learners and native English speakers: Contextualizing treatment impacts by classroom literacy instruction. *Journal of Educational Psychology*, **104**(4), 987–1005. doi:10.1037/a0028163

Vadasy, P., Sanders, E., & Abbott, R. (2008). Effects of supplemental early reading intervention at 2-year follow up: Reading skill growth patterns and predictors. *Scientific Studies of Reading*, **12**(1), 51–89. doi:10.1080/10888430701746906

Vadasy, P., Sanders, E., & Peyton, J. (2006). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology*, **98**(3), 508–528. doi:10.1037/0022-0663.98.3.508

Vadasy, P., Sanders, E., & Tudor, S. (2007). Effectiveness of paraeducator-supplemented individual instruction: Beyond basic decoding skills. *Journal of Learning Disabilities*, **40**(6), 508–525.

Vadasy, P., Wayne, S., O'Connor, R., Jenkins, J., Firebaugh, M., & Peyton, J. (2004). *Sound partners*. Denver: Sopris West.

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M., & Capin, P. (2016). Meta-analyses of the effects of Tier 2 type reading interventions in grades K–3. *Educational Psychology Review*, **28**, 551–571. doi:10.1007/s10648-015-9321-7

Weiser, B., & Mathes, P. (2011). Using encoding instruction to improve the reading and spelling performances of elementary students at risk for literacy difficulties: A best-evidence synthesis. *Review of Educational Research*, **81**(2), 170–200. doi:10.3102/0034654310396719

West, B., Welch, K., & Galecki, A. (2007). *Linear mixed models: A practical guide using statistical software*. Boca Raton, FL: CRC.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

WWC (What Works Clearinghouse). (2014). *Procedures and standards handbook* (Version 3.0). Washington, DC: Institute for Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf