Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Replicating the Impact of a Supplemental Beginning Reading Intervention: The Role of Instructional Context

Michael D. Coyne , Mary Little , D'Ann Rawlinson , Deborah Simmons , Oi-man Kwok , Minjun Kim , Leslie Simmons , Shanna Hagan-Burke & Christina Civetelli

To cite this article: Michael D. Coyne , Mary Little , D'Ann Rawlinson , Deborah Simmons , Oi-man Kwok , Minjun Kim , Leslie Simmons , Shanna Hagan-Burke & Christina Civetelli (2013) Replicating the Impact of a Supplemental Beginning Reading Intervention: The Role of Instructional Context, Journal of Research on Educational Effectiveness, 6:1, 1-23, DOI: 10.1080/19345747.2012.706694

To link to this article: https://doi.org/10.1080/19345747.2012.706694

Published online: 11 Jan 2013.

Journal of Research on Educational Effectiveness, 6: 1–23, 2013

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2012.706694

Routledge
Taylor & Francis Group

INTERVENTION, EVALUATION, AND POLICY STUDIES

Replicating the Impact of a Supplemental Beginning Reading Intervention: The Role of Instructional Context

Michael D. Coyne
University of Connecticut, Storrs, Connecticut, USA

Mary Little and D'Ann Rawlinson
University of Central Florida, Orlando, Florida, USA

Deborah Simmons
Texas A&M University, College Station, Texas, USA

Oi-man Kwok, Minjun Kim, Leslie Simmons, and Shanna Hagan-Burke
Texas A&M University, College Station, Texas, USA

Christina Civetelli
University of Connecticut, Storrs, Connecticut, USA

Abstract: The purpose of this varied replication study was to evaluate the effects of a supplemental reading intervention on the beginning reading performance of kindergarten students in a different geographical location and in a different instructional context from the initial randomized trial. A second purpose was to investigate whether students who received the intervention across both the initial and replication studies demonstrated similar learning outcomes. Kindergarten students ($n = 162$) identified as at risk of reading difficulty from 48 classrooms were assigned randomly at the classroom level either to a commercial program (i.e., Early Reading Intervention; Pearson/Scott Foresman, 2004) that included explicit/systematic instruction (experimental group) or school-designed typical practice intervention (comparison group). Both interventions were taught by classroom teachers for 30 min per day in small groups for approximately 100 sessions. Multilevel hierarchical linear analyses revealed no statistically significant differences between conditions on any measure. Combined analyses that included students from both the initial and replication studies suggested that differences in the impact of the intervention across studies were largely explained by mean differences in the comparison group students' response to school-designed intervention.

Keywords: Beginning reading, replication study, kindergarten students, experimental research, supplemental intervention

An important goal of increasing importance for reading researchers is to evaluate the efficacy of interventions through rigorous experimental research. The growing emphasis on

Address correspondence to Michael D. Coyne, University of Connecticut, Educational Psychology, 249 Glenbrook Rd, Unit 2064, Storrs, CT 06269, USA. E-mail: mike.coyne@uconn.edu

conducting randomized control trials in education has been influenced by policy documents (e.g., National Reading Panel, 2000; National Research Council, 2002), the priorities of Congress (Educational Sciences Reform Act of 2002), and governmental funding agencies (e.g., Institute of Educational Sciences). Findings from subsequent rigorously designed experimental research in reading have been disseminated widely in scholarly journals, in policy documents (e.g., National Early Literacy Panel, 2008), and by the U.S. Department of Education (e.g., What Works Clearinghouse, Institute of Educational Sciences Practice Guides) and support the positive benefits of beginning reading interventions for young students at risk of experiencing learning difficulties.

Evidence from an initial randomized controlled trial provides information about the impact of a particular beginning reading intervention under specific conditions and contexts. The goal of replication studies of educational interventions, in turn, is to determine whether the effects generalize to other settings under different conditions. To make informed decisions about beginning reading practices, practitioners and policymakers need evidence from multiple, high-quality studies evaluating the impact of interventions under different conditions (Gersten et al., 2005; What Works Clearinghouse, 2008). The purpose of this study was to conduct a rigorous varied replication evaluation of a kindergarten supplemental reading intervention in a different geographical location in schools characterized by a different instructional context than those in the initial randomized trial.

REPLICATION STUDIES OF BEGINNING READING INTERVENTIONS

Randomized control trials are designed to produce high-quality trustworthy evidence of effectiveness (Shadish, Cook, & Campbell, 2002; What Works Clearinghouse, 2008). However, the qualities that make the evidence from experimental studies trustworthy are the very same that may limit generalizability of findings from a single study (Fritz & Cleland, 2003). Specifically, research that emphasizes internal validity often has limited external validity. Randomized control trials are designed to ensure that effects can be attributed to a well-defined independent variable. At the same time, these studies reflect a single implementation under a somewhat unique set of contextual variables that do not always reflect realistic conditions across other schools and settings.

Replication studies, therefore, play a key role in strengthening the external validity and generalizibility of experimental research by evaluating whether the impact from initial studies can be reproduced with a separate set of participants across settings and instructional contexts. In educational research, replication studies of instructional practices are very rarely "true" replications that duplicate exactly the features and conditions of the initial study. Instead, in educational replication studies, researchers often seek to vary key features from earlier trials to investigate whether an intervention can produce similar effects under different conditions. These types of replications can be referred to as "varied" replications (Van Ijzendoorn, 1994), replications that "systematically vary one or more parameters of the original study to see whether its outcome remains stable or changes in a predictable way" (p. 57).

For example, studies may be conducted in schools and districts geographically distant from the initial sites. Studies may purposefully take place in schools that serve students with different demographics than in an earlier study; for example, students with diverse ethnicities, different socioeconomic status, or dissimilar achievement levels. Teachers and interventionists may also differ in experience, background, and familiarity with certain educational practices.

Key variables in intervention research are related to the instructional context of the schools and classrooms in which the intervention is being evaluated. For example, in studies

in which an intervention supplements typical practice, the type and quality of the general classroom instruction that is provided to all students is central. Another critical instructional variable in a varied replication study is the nature and quality of the instruction provided to the comparison group, especially if the intervention is compared to an alternative treatment rather than to a no treatment control. Experimental interventions are often compared to typical practice or "business as usual," which can vary considerably from school to school and even from year to year.

Evidence from varied replication and scale-up studies suggests that variability in the instructional context of schools may play a role in whether effects from earlier trials replicate. For example, recent research on Kindergarten Peer Assisted Learning Strategies (K-PALS; Fuchs, Fuchs, Thompson, Al Otaiba, Yen, McMaster, et al., 2001) underscored the importance of contextual and instructional factors in replication or scale-up studies. K-PALS, a classroom peer-tutoring early reading intervention, has been evaluated in a number of studies dating back to the mid-1990s, and overall evidence from early studies consistently found moderate to large treatment effects favoring students who received PALS over control students (Fuchs, Fuchs, Thompson, Al Otaiba, Yen, Yang, et al., 2001; Lemons, Fuchs, & Fuchs, 2008). However, a recent scale-up study of K-PALS in three states (Tennessee, Minnesota, Texas) found impacts that were more modest than in earlier studies (McMaster et al., 2010). After disaggregating the data, the researchers found that treatment effects differed considerably by geographical site with the largest effects in Tennessee and the smallest effects in Texas. When considering these findings, McMaster et al. concluded that schools in Tennessee had more instructional experience with K-PALS, were closer to the program developers, and had access to more resources whereas schools in Texas had the least amount of experience with K-PALS and also had more competing district demands.

In a follow-up investigation of the impact of K-PALS across studies, Lemons et al. (2008) evaluated the pretest to posttest gains in letter-sound knowledge, phonemic segmentation, word attack, and word identification made by both treatment and control students in different studies over time. Findings suggested that students who received K-PALS consistently made pre–post growth on all measures and that gains were larger in the more recent studies. Students in the control groups in these studies also made pre–post gains. Of interest, however, control group student gains increased more dramatically across studies and over time so that in the most recent studies, the overall impact of K-PALS, indexed by effect size differences between treatment and controls groups at posttest, was smaller even though students receiving the treatment experienced comparable or even larger gains than those in early studies that showed larger treatment effects. Lemons suggested that (a) "typical practice" in early literacy may be improving over time and (b) it is important to understand the quality of the instruction provided to the control group when interpreting treatment effects.

PURPOSE OF THE STUDY

The purpose of this study was to conduct a rigorous varied replication evaluation of a beginning reading intervention in schools characterized by a different instructional context than in the initial study. The Early Reading Intervention (ERI) is a commercially available kindergarten beginning reading intervention designed to supplement classroom instruction for students at risk of experiencing reading difficulties (Early Reading Intervention; Pearson/Scott Foresman, 2004). In our initial study (Simmons, Coyne, Hagan-Burke, Kwok, Simmons, Johnson et al., 2011), students ($N = 206$) were assigned randomly at the

classroom level to the experimental group or a comparison group (i.e., a school-designed intervention). Both interventions were taught for 30 min per day in small groups of three to five students for approximately 100 sessions. Multilevel hierarchical linear analyses revealed statistically significant effects favoring the ERI on foundational alphabetic, phonemic, and untimed decoding skills with effect sizes ranging from .40 to .51.

The initial study took place in six different school districts in Connecticut and Texas characterized by a distinctive instructional context. In these schools, core reading instruction varied widely both within and across classrooms and schools. Although some schools used basal programs, most schools used a combination of published materials and less structured guided-reading strategies. In addition, a majority of these schools did not routinely provide systematic supplemental kindergarten reading intervention; therefore implementing ERI represented a significant extension of their current reading practices. Overall, these districts included schools with a less coordinated and more individual approach to providing kindergarten reading instruction and intervention.

In the current study, we were interested in determining whether the impact of ERI would replicate in a school district in a different geographical region of the country that had a very different instructional context. Therefore, we evaluated ERI in a Florida school district that had a more coordinated, systematic, and consistent approach to providing beginning reading instruction and intervention. Kindergarten teachers in schools in this district had recently received professional development related to evidence-based reading instruction and delivering common core instruction using a published commercial program. This district also had experience providing small-group kindergarten reading intervention to students at risk for reading difficulties. We believed that evaluating ERI in schools in this district would extend the findings from our initial study and provide information about the efficacy of ERI in schools with a very different instructional context.

In summary, we were interested in whether the impact of ERI in our initial study would replicate in the Florida school district. We also wanted to know whether the effects of ERI on the absolute level of student performance measured at posttest would be similar across studies. Finally, we were interested in learning whether the posttest performance of students in the comparison groups, who received different school-designed interventions, would be comparable across studies.

METHOD

Initial Study of the ERI

Our initial randomized trial of the ERI, which was conducted 1 year prior to the replication study, is reported in detail in Simmons, Coyne, Hagan-Burke, Kwok, Simmons, Johnson et al. (2011). In the initial study, we worked with 206 kindergarten students in 57 classrooms from four schools in Texas and eight schools in Connecticut. The design, methodology, measures, and procedures in the initial study largely paralleled those in the replication study described next. Because we were interested in comparing the effects of ERI across studies, we include selected data and information from the first study as well as the replication study in Tables 1 to 11.

ERI Varied Replication Study

Schools and Interventionists. The varied replication study was conducted in a large school district in central Florida. School district administrators and researchers identified schools

Table 1. School and interventionist demographics

	Initial Study		Replication Study	
Variable	ERI	SDI	ERI	SDI
Schools	12		10	
Title I schools	11		8	
Interventionists	31	26	25	23
Interventionist experience				
Years teaching kindergarten (*M*)	6.24	7.92	4.47	8.40
SD	(7.00)	(8.89)	(3.90)	(7.29)

Note. ERI = Early Reading Intervention; SDI = school-designed intervention.

and facilitated cooperation among school administrators and kindergarten teachers. Schools represented a mixture of suburban and rural neighborhoods from the area. In total, 48 kindergarten teachers and 162 students from 10 elementary schools participated. Research staff met with the principals of each school to explain the study, the research design, and the need for randomization. All kindergarten classrooms from the 10 schools were then randomly assigned within schools to either treatment (ERI; $n = 25$) or comparison (school-designed intervention [SDI]; $n = 23$) conditions. Kindergarten classroom teachers served as interventionists in each classroom. Demographic data for schools and interventionists are found in Table 1.

Students. A two-phase process was used to identify children who were eligible to participate in the study as follows. In the 2nd month of kindergarten, researchers consulted with school personnel to identify children (a) who were considered in need of supplemental, small-group reading intervention; (b) who were at least 5 years of age; and (c) who received reading instruction in English. School personnel examined existing school-administered reading measures (e.g., Dynamic Indicators of Basic Early Literacy Skills [DIBELS], Good & Kaminski, 2002; etc.), consulted kindergarten teachers and nominated six to seven children per kindergarten classroom. Teachers were consulted to increase the probability that we identified children most at risk of reading difficulty. Permission forms were sent home with all nominated students. Of approximately 288 nominated students, 215 permission forms were returned. Demographic data for student participants may be found in Table 2

Next, nominated students with parental consent were screened using the (a) DIBELS Letter Naming Fluency (LNF; Good & Kaminski, 2002) and (b) Sound Matching subtest from the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen, & Rashotte, 1999). Screening measures were selected based on strong predictive validity for end-of-first-grade reading outcomes and prior use in recent kindergarten intervention studies (Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004).

To estimate risk, we identified scores on each measure that were as close to the 30th percentile as possible. Students were determined to be at risk and qualified to participate in the study if they met either of the following criteria: (a) a score at or below the 33rd percentile on the DIBELS LNF measure (i.e., fewer than six letters correctly named in 1 min) or (b) a score below the 37th percentile on the CTOPP Sound-Matching subtest. Up to five children per classroom were selected to participate.

Table 2. Student demographics

	Initial Study		Replication Study	
Variable	ERI[a]	SDI[b]	ERI[c]	SDI[d]
Gender				
Male	59 (52.7%)	47 (50.0%)	50 (57.5%)	38 (50.7%)
Female	53 (47.3%)	47 (50.0%)	37 (42.5%)	37 (49.3%)
Ethnicity				
Caucasian	42 (37.5%)	37 (39.4%)	52 (59.8%)	46 (61.3%)
Latino	44 (39.3%)	42 (44.7%)	23 (26.4%)	16 (21.3%)
African American	24 (21.4%)	15 (16.0%)	5 (5.7%)	10 (13.3%)
Other	2 (1.8%)	—	7 (8.0%)	3 (4.0%)
English Language Learner	24 (21.4%)	29 (30.9%)	15 (17.2%)	10(13.3%)
Age (M)	5.49	5.42	5.53	5.53
SD	(0.32)	(0.27)	(0.33)	(0.33)

Note. No differences were statistically significant. ERI = Early Reading Intervention; SDI = school-designed intervention.
[a]$N = 112$. [b]$N = 94$. [c]$N = 87$. [d]$N = 75$.

PROCEDURES

Core Reading Instruction

All schools in this Florida school district implemented a consistent core reading program (*Harcourt Trophies*; Beck, Farr, & Strickland, 2007). Harcourt Trophies is a comprehensive reading/language arts program that uses explicit phonics instruction, guided reading strategies, phonemic awareness instruction, systematic intervention strategies, integrated language arts components, and aligned assessment tools. Teachers in the participating school district received systematic professional development in implementing the core reading program. By state and district policies, all kindergarten students receive district-prescribed core reading instruction for 90 min each school day.

Common Intervention Components

To increase comparability between conditions, a number of common instructional components were standardized across both the experimental ERI and the comparison SDI conditions. Groups in both conditions comprised three to five students. Interventionists in both conditions were asked to meet with their groups for 30 min 5 days per week over the course of the intervention period for an equivalent number of total sessions. Content in both conditions focused on early literacy skills, with ERI interventionists implementing the ERI curriculum and comparison interventionists providing school-designed reading intervention focusing on early literacy skills.

ERI Intervention

The ERI curriculum includes 126 daily lessons. A 30-min lesson consists of seven activities, each designed to last 3 to 5 min. The first 15 min of the lesson focus on phonological awareness and alphabetic understanding; the second 15 min integrate writing and spelling with previously taught phonemic and alphabetic skills.

The program is organized into four major components. Part I: Learning Letters and Sounds consists of 42 lessons and introduces 11 letter names and sounds and the phonemic skills of first and last sound isolation. Part II: Segmenting, Blending, and Integrating includes 30 lessons and continues with the introduction of five new letter names and sounds while introducing phonemic blending and segmenting using letter tiles. Part III: Reading Words completes the introduction of the six remaining letter names and sounds with primary instruction focusing on word decoding in vowel-consonant and consonant-vowel-consonant words. This part consists of 24 lessons in which instruction integrates oral segmenting and blending with real-word decoding and the introduction of irregular word reading. Part IV: Reading Sentences and Storybooks consists of 30 lessons. Instruction in this final section focuses on combining alphabetic skills and strategies with irregular word reading to read sentences and short storybooks.

The 126 lessons are highly specified and include detailed scripting to ensure clear and consistent communication of information and reduce variability in implementation. When introducing a new skill, the teacher models the information several times using consistent wording. In addition, skills are carefully integrated to enhance learning. Scheduled instruction, review, and feedback are explicitly incorporated in the program. For example, each lesson that introduces new information includes a specified number of instructional interactions in which the teacher first models the information. Students practice the new skill with the teacher and then apply it to new untaught discrimination or generalization tasks. In addition, the intervention provides teachers explicit instructional language and procedures for correcting errors and extending practice for difficult items.

ERI Training. Professional development was designed to approximate the type of training that is typical when schools adopt published programs; it was limited to 2 days. The 1st day of training took place before the start of the intervention. The goals of this professional development session were to develop an awareness of the purpose of the research project, to orient the interventionists to the design of the ERI program, to introduce the ERI curriculum and materials, and to provide guidance for implementation of the intervention. The professional development familiarized interventionists with lesson structure and the scope and sequence of Parts 1 and 2 of the four-part ERI curriculum. Interventionists viewed publisher-developed video clips of lesson elements and participated in hands-on sessions with the curriculum materials. They were also shown how to set up and manage materials and student groups. Finally, time was devoted to addressing critical instructional techniques such as giving immediate corrective feedback and providing both group and individual turns to students.

The 2nd day of professional development was held after the interventionists had completed half of the ERI curriculum. This took place in late January. This component of the professional development focused on the lessons and materials for Parts 3 and 4 of the program and followed a format similar to that of the first professional development day. Project ERI researcher team members led both professional development sessions. The same researchers also provided professional development training during the initial study. ERI training was comparable across the initial and replication studies.

SDI

Kindergarten teachers who served as interventionists in the comparison, or SDI, condition were asked to provide typical school-designed beginning reading intervention to identified students for 30 min daily, in groups of three to five students. Although these interventionists did not receive any additional professional development, all kindergarten teachers in

this school district were experienced in providing supplemental beginning reading intervention and had previously received extensive professional development and resources in evidence-based reading instruction methods, both at the district and school site, through the use of literacy coaches. Typical kindergarten intervention supports in this school district include strategically integrated phoneme awareness and alphabetic understanding instruction. During the year of this study, one 120-min professional development session was provided by school district reading coaches using K-1 Student Center Activity resources from the Florida Center for Reading Research (Florida Center for Reading Research, 2005), curriculum-based formative assessment data, and individual student achievement data in reading.

Fidelity of Implementation

Our approach to documenting and measuring fidelity of implementation addressed four distinct dimensions of treatment integrity (Dane & Schneider, 1998; Gersten et al., 2005): (a) procedural fidelity and adherence to the ERI program, (b) quality of the instructional delivery of both the ERI and SDI interventions, (c) dosage of the ERI and SDI interventions, and (d) documentation of the content focus of the SDI intervention to evaluate program differentiation compared to ERI.

Procedural fidelity and adherence of the ERI intervention was assessed by direct observations using a checklist developed to document the presence or absence of key instructional features of the intervention (Gresham, MacMillan, Beebe-Frankenberger, & Bocian, 2000). For each of the seven activities within an ERI lesson, observers completed four items evaluating whether interventionists completed each component of the activity and whether they were fluent with lesson wording and activities. Procedural fidelity items were evaluated on a 1-to-4 scale, with 4 indicating excellent, 3 indicating good, 2 indicating fair, and 1 indicating poor levels of fidelity.

Quality of the instructional delivery of the ERI and SDI interventions was also assessed through direct observations. Three observation items focused on quality features such as the interventionist's ability to manage instructional time, maximize students' opportunities to practice, and provide corrective feedback. Quality fidelity items were evaluated on the same 1-to-4 scale. Research staff observed ERI and SDI interventionists three times (i.e., fall, winter, spring) over the course of the school year.

Dosage of the ERI and SDI conditions was documented through interventionist logs, in which teachers recorded the date, the lesson number, and the students who were present for the intervention. We also evaluated program differentiation within the comparison condition, or the extent to which the SDI instruction differed across the two studies. Here observers recorded the instructional content of each activity within the SDI lessons observed from a list of 13 early literacy areas. If an activity focused on two different content areas (e.g., phonemic awareness and letter-sound knowledge), both were recorded. These procedures allowed us to determine the percentage of SDI lessons that included an emphasis on specific areas of beginning reading instruction.

Measures and Assessment Procedures

The assessment battery included screening, pretest, and posttest assessments. Assessments measured a range of early literacy and beginning reading constructs. Screening and pretest

assessments were completed in September prior to the start of the intervention. Posttesting occurred within 2 weeks after the completion of the intervention in May.

Letter Knowledge. The LNF subtest of the DIBELS (Good & Kaminski, 2002) was used as a screening measure. The LNF measures a student's ability to rapidly name (in 1 min) upper- and lowercase letters presented. Alternate-form reliability for the kindergarten sample is .89. The split-half reliability coefficient for this subtest is .94 for the first-grade sample (coefficient for kindergarten sample not reported). The Supplementary Letter Checklist of the Woodcock Reading Mastery Tests–Revised/Normative Update (WRMT-R/NU; Woodcock, 1987, 1998) was administered at posttest to assess children's ability to correctly identify the letter name and produce the appropriate sound for each presented lowercase letter.

Phonological Awareness. The Sound-Matching subtest of the CTOPP (Wagner et al., 1999) measures a child's ability to select one of three pictures that has the same initial or final sound as the first word. This is an untimed measure. The internal consistency coefficient (Cronbach's alpha) is .93 for both the age 5 and age 6 samples. The Sound-Matching subtest was used as a screening measure and as well as the posttest. The Blending Words subtest of the CTOPP (Wagner et al.) is an untimed measure that requires the student to combine sounds into real words. Sounds are presented on an audiocassette tape. The internal consistency coefficient (Cronbach's alpha) is .88 for the age 5 sample and .89 for the age 6 sample. The Phoneme Segmentation Fluency DIBELS subtest (Good & Kaminski, 2002) measures a student's ability to fluently segment three- and four-phoneme words. Scores indicate the number of sound segments correctly identified in 1 min. Alternate-form reliability for the kindergarten sample is .88. The Blending Words and the Phoneme Segmentation Fluency measures were administered at both pre- and posttest.

Decoding. The Word Attack subtest of the WRMT-R/NU (Woodcock, 1987, 1998) is an untimed measure of a student's skill in reading a list of nonwords (e.g., "tet") presented in isolation. The raw score is the number of nonwords read correctly, which is converted into a standard score. The split-half reliability coefficient for this subtest is .94 for the first-grade sample (coefficient for kindergarten sample not reported). The Nonsense Word Fluency (NWF) subtest of the DIBELS (Good & Kaminski, 2002) measures a student's knowledge of letter-sound correspondence and the ability to fluently blend letters into words in which letters represent their most common sounds. Scores indicate the number of letter sounds produced correctly in 1 min. Because the measure is fluency based, students receive a higher score if they read the nonsense word as a whole word and a lower score if they give letter sounds in isolation. The alternate-form reliability for this subtest for the kindergarten sample is .88. The Word Attack and Nonsense Word Fluency measures were administered at both pre- and posttest.

Word Identification. The Word Identification subtest of the WRMT-R/NU (Woodcock, 1987, 1998) measures a student's skill in reading a list of real words presented in isolation and is untimed. The split-half reliability coefficient is .98 for the first-grade sample (coefficient for kindergarten sample not reported). The Word ID measure was administered at both pre- and posttest.

Vocabulary. Vocabulary was assessed at pretest only, using the Peabody Picture Vocabulary Test–Third Edition (PPVT–III; Dunn & Dunn, 1997), an individually administered oral test

of receptive vocabulary. For each test item, the student is presented with four black-and-white illustrations and asked to select the picture considered to best illustrate the meaning of a word presented orally by the examiner. The internal-consistency reliability (Cronbach's alpha) for form IIIA is .95 for the 5- to 6-year-old group, and the test–retest reliability ranges from .92 to .93.

Procedures. All assessments were administered by trained data collectors individually to students in quiet locations outside of the classroom. Data collectors included graduate students who had participated in training that consisted of a review of general assessment procedures, modeling of the specific test protocols, paired practice, and supervised independent practice of each test. Each data collector met the criterion of 90% accuracy in recording scores for a modeled administration of each measure. After data collection, each testing protocol was independently scored by two trained individuals.

RESULTS

Data Analysis Approach

To evaluate the impact of the ERI intervention in the varied replication study and compare the effects of ERI across the initial and replication studies, we included data from both studies in one multiyear model. We looked at four different comparisons across four different groups of students: (a) students who received the ERI intervention in the initial study compared to students who received the comparison intervention in the initial study, (b) students who received the ERI intervention in the replication study compared to students who received the comparison intervention in the replication study, (c) students who received the ERI intervention in the initial study compared to students who received the ERI intervention in the replication study, and (d) students who received the comparison intervention in the initial study compared to students who received the comparison intervention in the replication study. It is important to note that only Comparisons a and b are experimental (i.e., groups randomized). Comparisons c and d were included to examine whether performance of students in the ERI and comparison groups were similar or different across the initial and replication studies. However, because these comparisons were not experimental (i.e., groups were preexisting), results should be interpreted with more caution. Furthermore, we use effect size terminology to interpret the magnitude of differences among all group comparisons. Again, because Comparisons c and d are not experimental, values should be interpreted as standardized differences presented as effect sizes instead of true effect sizes.

Because of the nested structure of our data (i.e., students nested within intervention groups nested within schools), multilevel modeling (Hox, 2002) was chosen to analyze the data using Hierarchical Linear Model (HLM; V6.08; Raudenbush, Bryk, Cheong, & Congdon, 2004). The three-level analysis included students (i.e., Level 1) nested within interventionists (i.e., Level 2) and further nested within schools (i.e., Level 3). All models included the following covariates: student pretest scores (matched pretest when available or WRMT-R/NU Letter Identification pretest standard score), PPVT–III pretest standard score, gender, age, ethnicity, special education status, and English language learner status. Full Information Maximum likelihood was used to estimate all models, and two-tailed tests along with Hedges's effect size (δ_w; Hedges, 2007) were used to evaluate the potential group differences. In addition, the Benjamini–Hochberg correction (Benjamini &

Hochberg, 1995) was adopted for controlling for the comparison-wise type I error rate as recommended by the What Works Clearinghouse (2008).

Pretest Analyses

Descriptive statistics of pretest measures for the ERI and SDI conditions for both the initial and the replication study are presented in Table 3. Analysis of *t* tests revealed no statistically significant differences between any of the four groups (i.e., ERI/SDI; Year 1/Year 2) on any measure. Chi-square analyses used to test group differences on student demographic variables (i.e., gender, ethnicity, and English language learner status) indicated no statistically significant differences between groups on any variables.

Attrition Analyses

In this (FL) replication study, 188 kindergarten students were selected to participate. Of this group, 162 (86.2%) participated in both pretest and posttest assessments; 26 (13.8%) did not complete the study and were not available at posttest. In the initial (Texas and Connecticut) study, 232 kindergarten students were selected to participate. Of this group, 206 (88.8%) participated in both pretest and posttest assessments; 26 (11.2%) did not complete the study and were not available at posttest. We conducted chi-square tests to determine any differences in attrition patterns for each of the 2 years between students who exited the study and those who remained. There was no statistically significant difference in the attrition rate between the ERI group and the SDI group across the two studies.

Fidelity of Implementation

Table 4 displays fidelity data for both the initial and the replication study. Procedural fidelity items were evaluated on a 1-to-4 scale, with 4 indicating excellent, 3 indicating good, 2 indicating fair, and 1 indicating poor levels of fidelity. Mean procedural fidelity for the ERI interventionists was 3.07 for the initial study and 2.84 for the replication study, suggesting that interventionists implemented ERI with generally satisfactory levels of integrity. Items related to the quality of implementation of both ERI and SDI was also evaluated on a 1-to-4 scale. Mean quality scores for interventionists across studies and instructional conditions ranged from 2.52 to 3.01, indicating that the quality of instruction provided to students was fair to good. Dosage refers to the average number of lessons provided to students in each condition across years. Dosage data indicated that, on average, students received similar numbers of lessons.

Table 5 provides information about the percentage of SDI lessons observed that included specific areas of beginning reading instruction. These data allowed us to investigate program differentiation (Dane & Schneider, 1998), or the extent to which the SDI instruction differed across the two studies. Findings suggest that although the general content focus was similar in the SDI condition during the initial and the replication study, there were some clear differences in instructional emphasis across the two studies. That is, the SDI instruction during the replication study included a greater emphasis on the phonological skills of blending and segmenting, sight word work, reading connected text, and writing sounds and words.

Table 3. Descriptive statistics for student measures at pretest and posttest

	Initial Study				Replication Study			
	Pretest		Posttest		Pretest		Posttest	
Measure	ERI *M (SD)*	SDI *M (SD)*	ERI *M (SD)*	SDI *M (SD)*	ERI *M (SD)*	SDI *M (SD)*	ERI*M (SD)*	SDI *M (SD)*
Alphabet Knowledge[b]								
Letter Identification[b]	87.04 (10.2)	88.09 (11.52)			89.96 (10.13)	90.14 (10.13)		
WRMT-R/NU Supplementary[b] Letter Checklist-Name[a]			25.61 (4.28)	25.00 (4.80)			25.87 (3.85)	26.15 (2.89)
Letter Sound Knowledge[b]								
WRMT-R/NU Supplementary[b] Letter Checklist-Sounds[a]			24.93 (6.00)	22.86 (7.50)			26.15 (6.22)	25.76 (6.20)
Phonemic Awareness[b]								
CTOPP Sound Matching[b]	7.68 (1.10)	7.69 (1.15)	9.64 (1.97)	9.05 (2.06)	7.60 (1.12)	7.73 (1.23)	9.61 (1.92)	9.37 (1.92)
CTOPP Blending Words[b]	8.52 (1.66)	8.46 (1.80)	10.49 (1.96)	10.00 (2.33)	7.92 (1.97)	8.43 (2.03)	10.87 (2.49)	11.27 (2.25)
DIBELS Phonemic[b] Segmentation Fluency[a]			32.48 (15.07)	26.20 (18.93)			43.86 (18.15)	44.51 (16.32)
Word Attack[b]								
WRMT-R/NU Word Attack[b]	94.73 (2.66)	94.78 (3.12)	108.09 (9.20)	105.16 (10.99)	94.11 (1.07)	94.61 (3.77)	108.09 (10.77)	108.60 (10.82)
DIBELS Nonsense Word[b] Fluency[a]			25.74 (11.67)	22.40 (14.40)			30.08 (14.39)	32.71 (15.87)
Word Identification[b]								
WRMT-R/NU Word ID[b]	83.56 (6.71)	84.29 (8.17)	104.10 (11.15)	103.26 (14.75)	83.82 (7.49)	84.79 (9.26)	107.13 (12.00)	109.19 (10.95)
Vocabulary[b]								
Peabody Picture Vocabulary Test[b]	87.55 (15.03)	88.74 (15.33)			90.55 (13.44)	92.42 (11.90)		

Note. No pretest differences were statistically significant. ERI = Early Reading Intervention; SDI = school-designed intervention; WRMT-R/NU = Woodcock Reading Mastery Tests–Revised/Normative Update; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

[a]Raw score. [b]Standard score.

Table 4. Fidelity of implementation data

	Initial Study		Replication Study	
Variable	ERI *M (SD)*	SDI *M (SD)*	ERI *M (SD)*	SDI *M (SD)*
Procedural fidelity	3.07 (0.36)	n/a	2.84 (0.43)	n/a
Quality	3.01 (0.38)	2.64 (0.42)	2.52 (0.41)	2.63 (0.34)
Dosage	102.47 (10.10)	106.03 (14.87)	97.77 (14.91)	87.69 (23.27)

Note. Quality = Quality of instructional delivery; Dosage = Number of days intervention was delivered. Fidelity ratings: 4 = excellent, 3 = good, 2 = fair, and 1 = poor. ERI = Early Reading Intervention; SDI = school-designed intervention.

Posttest Analyses

Descriptive statistics for posttest measures for both the initial and replication studies are included in Table 3. We further examined the posttest mean differences between conditions using HLM. Four models were analyzed: (a) 1st-year ERI ($n = 112$) versus 1st-year SDI ($n = 94$), (b) 2nd-year ERI ($n = 87$) versus 2nd-year SDI ($n = 75$), (c) 1st-year ERI versus 2nd-year ERI, and (d) 1st-year SDI versus 2nd-year SDI. HLM analyses were conducted with the following set of models:

Level 1 (student-level) model

$$\begin{aligned}\text{Posttest}_{ijk} = {} & \pi_{0jk} + \pi_{1jk}\text{Pretest}_{ijk} + \pi_{2jk}\text{PPVT}_{ijk} + \pi_{3jk}\text{Age}_{ijk} + \pi_{4jk}\text{Gender}_{ijk} \\ & + \pi_{5jk}\text{Hispanic}_{ijk} + \pi_{6jk}\text{African\_American}_{ijk} + \pi_{7jk}\text{Special\_ed}_{ijk} \\ & + \pi_{8jk}\text{Bilingual}_{ijk} + e_{ijk} \qquad (1)\end{aligned}$$

Table 5. Percentage of school-designed intervention lessons observed that included an emphasis on specific content

Content	Initial Study	Replication Study
Phonological awareness		
Word/syllable level	13.7	3.7
Phoneme level	41.2	46.3
Blending/segmenting	19.6	35.2
Alphabetics/Phonics		
Letter names	66.7	64.8
Letter sounds	62.8	72.2
Reading words	41.2	48.1
Sight word work	19.6	42.6
Reading connected text	27.5	46.3
Writing		
Letter names	23.5	20.4
Letter sounds	11.8	38.9
Words	13.7	37.0
Vocabulary	13.7	7.4
Listening comprehension	13.7	7.4

where i represents the i-th student, j represents the j-th group, and k represents the k-th school. Posttest_{ijk} is the score of one of the posttest measures for the i-th student from the j-th group in the k-th school.

In this student-level model, covariates included the corresponding pretest score (Pretest_{ijk}), the PPVT score (PPVT_{ijk}) as an indicator of the student's receptive language ability, plus demographic variables consisting of student age (Age_{ijk}), gender (Gender_{ijk}), ethnicity (represented by two dummy-coded variables with Caucasian students as the reference group: Hispanic_{ijk} and $\text{African American}_{ijk}$), special education services (Special_ed_{ijk}), and bilingual status (Bilingual_{ijk}). Both the corresponding pretest score and the PPVT score were centered at the grand mean.[1] The within-group random error is e_{ijk}, and the corresponding variance, $V(e_{ijk}) = \sigma^2$, captures the within-group variation. For posttests for which there were no corresponding pretests, we examined correlations to identify the pretest most highly associated with posttest performance. The untimed letter identification pretest score was highly correlated with several posttest-only measures, including WRMT supplementary letter checklist name and sound, nonsense word fluency, and phonemic segmentation fluency test, and was used as the pretest covariate for these measures.

The intervention group-level models were specified as shown next:

Level 2 (group-level) models

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}\,\text{Intervention}_{jk} + r_{0jk};$$

and

$$\begin{aligned}\pi_{1jk} = \beta_{10k}; \pi_{2jk} = \beta_{20k}; \pi_{3jk} = \beta_{30k}; \pi_{4jk} = \beta_{40k}; \pi_{5jk} = \beta_{50k}; \pi_{6jk} = \beta_{60k}; \pi_{7jk} \\ = \beta_{70k}; \pi_{8jk} = \beta_{80k}\end{aligned} \tag{2}$$

Intervention_{jk} is a dummy-coded variable that examines the potential group difference in the four comparison models. That is, in Model 1, β_{10k} examines the intervention effect between 1st-year ERI and 1st-year SDI, whereas the same coefficient examines the intervention effect between 2nd-year ERI and 2nd-year SDI in Model 2. Similarly, β_{10k} examines the plausible group difference between 1st-year ERI and 2nd-year ERI in Model 3, and the same coefficient examines the plausible group difference between 1st-year SDI and 2nd-year SDI in Model 4. The between-group random effect is r_{0jk}, and the corresponding variance, $V(r_{0jk}) = \tau_{00}$, captures the between-group variation.

The school-level models were specified as shown next:

Level 3 (school-level) models

$$\beta_{00k} = \gamma_{000} + U_{00k};$$

and

$$\begin{aligned}\beta_{01k} = \gamma_{010}; \beta_{02k} = \gamma_{020}; \beta_{10k} = \gamma_{100}; \beta_{20k} = \gamma_{200}; \beta_{30k} \\ = \gamma_{300}; \beta_{40k} = \gamma_{400}; \beta_{50k} = \gamma_{500}; \beta_{60k} = \gamma_{600}; \beta_{70k} = \gamma_{700}; \beta_{80k} = \gamma_{800}.\end{aligned} \tag{3}$$

[1]Multivariate Linear Model analyses with group mean centering were conducted as well and showed similar results in terms of the statistical significance and effects.

Table 6. Tests of statistical significance between ERI and SDI groups for initial study

Measure	γ_{010}	σ^2	τ_{00}	τ_{000}	*Two-tailed p*
Alphabet Knowledge					
WRMT-R/NU Letter Name Checklist	0.76	15.76	1.70	2.97	0.182
Letter Sound Knowledge					
WRMT-R/NU Letter Sound Checklist	2.32	27.63	12.65	6.41	0.036
Phonemic Awareness					
CTOPP Sound Matching	0.77	3.25	0.53	0.31	0.009
CTOPP Blending Words	0.76	3.50	0.38	0.60	0.011
DIBELS Phonemic SegmentationFluency	6.69	214.86	47.43	30.62	0.012
Word Attack					
DIBELS Nonsense Word Fluency	3.87	113.56	37.89	20.63	0.070
WRMT-R/NU Word Attack	4.44	76.35	12.86	14.51	0.005
Word Identification					
WRMT-R/NU Word Identification (ID)	2.43	96.08	30.50	33.42	0.204

Note. School-designed intervention (SDI) is the reference group. Variance components (i.e., σ^2, τ_{00}, & τ_{000}) were based on the unconditional model.γ_{010}, as shown in equation (3), is the difference between Early Reading Intervention (ERI) and SDI groups for initial study on the posttest measure while holding the effects of other variables as constant. Positive value indicated that ERI group, on average, scored higher on the posttest measure than the comparison group. WRMT-R/NU = Woodcock Reading Mastery Tests–Revised/Normative Update; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.
[a]The approximate degrees of freedom (from HLM) for all the *t* tests are equal to 55.

The target effect, γ_{010}, represents the average group difference between different conditions (e.g., 1st-year ERI and 1st-year SDI) within each model after controlling for all the other variables, including the corresponding pretest covariate, demographic variables, and the teacher and school effects. The between-school random effect is U_{00k}, and the corresponding variance, $V(U_{00k}) = \tau_{000}$, captures the between-school variation.

Two intraclass correlations (ICCs), specifically, group-level ICC (i.e., $ICC_{group} = \rho* = \frac{\tau_{000}+\tau_{00}}{\tau_{000}+\tau_{00}+\sigma^2}$) and school-level ICC (i.e., $ICC_{school} = \frac{\tau_{000}+\tau_{00}}{\tau_{000}+\tau_{00}+\sigma^2}$), were calculated based on Hox's (2002) equations. The group-level ICCs ranged from .16 to .32, whereas the school-level ICCs ranged from .06 to .22. All these ICCs are in a range commonly seen in educational studies. To evaluate the significant group difference between different conditions, we used the Hedges's δ_w effect size.

Results of the four models are provided in Tables 6 to 9. Effect size results indicating standardized differences are presented in Table 10 as well as represented in Figure 1. The reference group for each comparison is denoted with an asterisk for each model. All statistically significant effects, after controlling for the comparison-wise type I error rate, are indicated in bold. The second column of Table 10 shows the effect size of the ERI condition compared to the SDI condition for the initial study. Statistically significant findings were found on the WRMT-R/NU Supplementary letter checklist sounds ($\delta = .44$), CTOPP sound matching ($\delta = .43$), CTOPP blending words ($\delta = .40$), DIBELS phonemic segmentation fluency ($\delta = .46$), and WRMT-R/NU word attack ($\delta = .51$), which all consistently favored the ERI condition over the 1st-year SDI condition.

The second column of Table 10 indicates that there were no statistically significant differences between the ERI and SDI conditions on any measures in the 2nd-year replication study after adjusting for the pretest covariate, the demographic variables, and the teacher

Table 7. Tests of statistical significance between ERI and SDI groups for replication study

Measure	γ_{010}	σ^2	τ_{00}	τ_{000}	*Two-tailed p*
Alphabet Knowledge					
WRMT-R/NU Letter Name Checklist	–0.33	9.07	1.11	1.65	0.474
Letter Sound Knowledge					
WRMT-R/NU Letter Sound Checklist	–0.20	0.05	5.86	–0.02	0.794
Phonemic Awareness					
CTOPP Sound Matching	0.27	3.21	0.26	0.17	0.294
CTOPP Blending Words	–0.16	4.81	0.00	0.85	0.631
DIBELS Phonemic Segmentation Fluency	–1.23	262.74	0.20	32.66	0.609
Word Attack					
DIBELS Nonsense Word Fluency	–2.91	193.57	0.30	34.45	0.166
WRMT-R/NU Word Attack	–0.03	104.33	0.09	9.97	0.986
Word Identification					
WRMT-R/NU Word Identification (ID)	–1.89	0.18	15.95	–0.12	0.241

Note. School-designed intervention (SDI) is the reference group. Variance components (i.e., σ^2, τ_{00}, & τ_{000}) were based on the unconditional model. γ_{001}, as shown in Equation 3, is the difference between Early Reading Intervention (ERI) and SDI groups for replication study on the posttest measure while holding the effects of other variables as constant. Positive value indicated that ERI group, on average, scored higher on the posttest measure than the comparison group. WRMT-R/NU = Woodcock Reading Mastery Tests–Revised/Normative Update; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.
[a]The approximate degrees of freedom (from HLM) for all the *t* tests are equal to 46.

Table 8. Tests of statistical significance between initial and replication study within ERI groups

Measure	γ_{010}	σ^2	τ_{00}	τ_{000}	*Two-tailed p*
Alphabet Knowledge					
WRMT-R/NU Letter Name Checklist	–0.63	14.05	1.63	1.12	0.351
Letter Sound Knowledge					
WRMT-R/NU Letter Sound Checklist	–0.11	28.89	2.89	5.96	0.929
Phonemic Awareness					
CTOPP Sound Matching	–0.17	3.31	0.00	0.51	0.616
CTOPP Blending Words	0.39	4.29	0.00	0.58	0.365
DIBELS Phonemic Segmentation Fluency	7.54	229.64	0.24	72.31	0.042
Word Attack					
DIBELS Nonsense Word Fluency	3.00	142.77	0.10	31.62	0.296
WRMT-R/NU Word Attack	–0.44	87.86	0.05	9.20	0.822
Word Identification					
WRMT-R/NU Word Identification (ID)	2.30	107.02	0.08	25.14	0.374

Note. Initial Early Reading Intervention (ERI) is the reference group. Variance components (i.e., σ^2, τ_{00}, & τ_{000}) were based on the unconditional model. γ_{010}, as shown in Equation 3, is the difference between initial and replication study within ERI groups on the posttest measure while holding the effects of other variables as constant. Positive value indicated that ERI groups in the replication study, on average, scored higher on the posttest measure than the comparison group. WRMT-R/NU = Woodcock Reading Mastery Tests–Revised/Normative Update; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.
[a]The approximate degrees of freedom (from HLM) for all the *t* tests are equal to 54.

Table 9. Tests of statistical significance between initial and replication study within SDI groups

Measure	γ_{010}	σ^2	τ_{00}	τ_{000}	*Two-tailed p*
Alphabet Knowledge					
WRMT-R/NU Letter Name Checklist	0.82	11.73	1.32	3.43	0.345
Letter Sound Knowledge					
WRMT-R/NU Letter Sound Checklist	2.51	31.66	13.61	4.15	0.049
Phonemic Awareness					
CTOPP Sound Matching	0.53	3.02	0.92	0.00	0.143
CTOPP Blending Words	1.61	3.99	0.37	1.11	0.003
DIBELS Phonemic Segmentation Fluency	16.35	237.54	37.62	100.10	0.000
Word Attack					
DIBELS Nonsense Word Fluency	9.68	159.81	33.78	58.49	0.002
WRMT-R/NU Word Attack	4.65	84.00	22.34	14.30	0.081
Word Identification					
WRMT-R/NU Word Identification (ID)	7.86	104.10	39.97	30.86	0.012

Note. Initial school-designed intervention (SDI) is the reference group. Variance components (i.e., σ^2, τ_{00}, & τ_{000}) were based on the unconditional model. γ_{001}, as shown in Equation 3, is the difference between initial and replication study within SDI groups on the posttest measure while holding the effects of other variables as constant. Positive value indicated that SDI groups in replication study, on average, scored higher on the posttest measure than the comparison group. WRMT-R/NU = Woodcock Reading Mastery Tests–Revised/Normative Update; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.
[a]The approximate degrees of freedom (from HLM) for all the *t* tests are equal to 47.

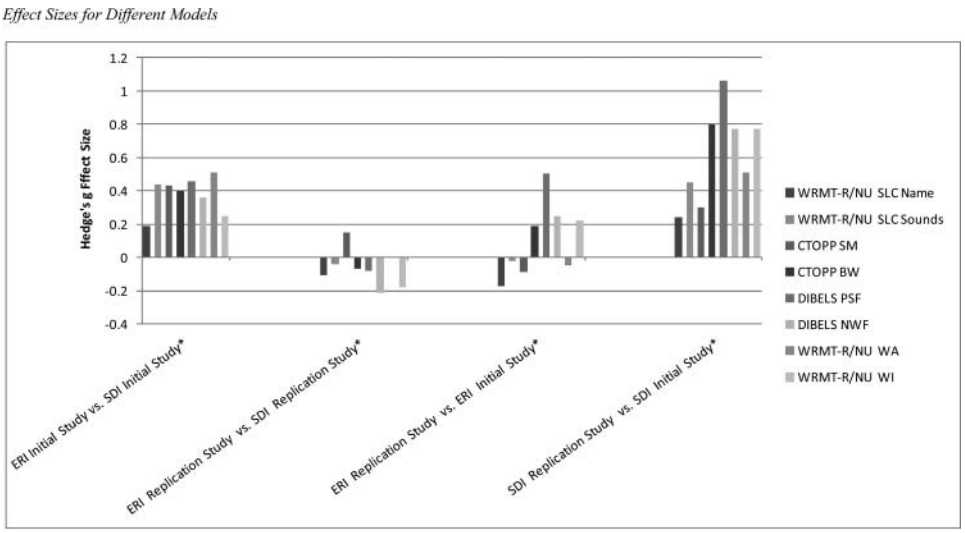

Figure 1. Effect sizes for different models. *Note*. *Reference group. WRMT-R/NU = Woodcock Reading Mastery Tests–Revised/Normative Update; SLC = Supplementary Letter Checklist; CTOPP = Comprehensive Test of Phonological Processing; SM = Sound-Matching; BW = Blending Words; DIBELS = Dynamic Indicators of Basic Early Literacy Skills; PSF = Phoneme Segmentation Fluency; NWF = Nonsense Word Fluency; WA = Word Attack; WI = Word Identification; ERI = Early Reading Intervention; SDI = school-designed intervention.

Table 10. Standardized differences (Hedges's *g*) for different models

Measure	ERI Initial Study vs. SDI Initial Study[a]	ERI Replication Study vs. SDI Replication Study[a]	ERI Replication Study vs. ERI Initial Study[a]	SDI Replication Study vs. SDI Initial Study[a]
Alphabet Knowledge				
WRMT-R/NU Supplementary Letter Checklist-Name	0.19	–0.11	–0.17	0.24
Letter Sound Knowledge				
WRMT-R/NU Supplementary Letter Checklist-Sounds	**0.44**	–0.04	–0.02	**0.45**
Phonemic Awareness				
CTOPP Sound Matching	**0.43**	0.15	–0.09	0.30
CTOPP Blending Words	**0.40**	–0.07	0.19	**0.80**
DIBELS Phonemic Segmentation	**0.46**	–0.08	**0.50**	**1.06**
Fluency				
Word Attack				
DIBELS Nonsense Word Fluency	0.36	–0.21	0.25	**0.77**
WRMT-R/NU Word Attack	**0.51**	0	–0.05	0.51
Word Identification				
WRMT-R/NU Word ID	0.25	–0.18	0.22	**0.77**

Note. Bold indicates a significant effect. ERI = Early Reading Intervention; SDI = school-designed intervention; WRMT-R/NU = Woodcock Reading Mastery Tests–Revised/Normative Update; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Early Literacy Skills.
[a]Reference group.

and school effects. The next two columns show the standardized effect sizes comparing performances by year for each intervention condition. In general, the 2nd-year ERI group (for the replication study) was very similar to the 1st-year ERI group (for the initial study) on all outcome measures except the DIBELS phoneme segmentation fluency, in which the 2nd-year ERI group outperformed the 1st-year ERI group ($\delta = .50$). However, the 2nd-year SDI group (replication study) performed significantly better than the 1st-year SDI group (initial study) on several outcome measures, including the WRMT-R/NU Supplementary letter checklist sounds ($\delta = .45$), CTOPP blending words ($\delta = .80$), DIBELS phonemic segmentation fluency ($\delta = 1.06$), DIBELS nonsense word fluency ($\delta = .77$), and WRMT-R/NU word identification ($\delta = .77$).

DISCUSSION

Summary of Findings

In this study, we were interested in determining whether the positive impact of a supplemental kindergarten reading intervention (i.e., the Early Reading Intervention) in an initial randomized study would replicate in a different school district with a different instructional

context 1 year later. We also wanted to know whether the effects of the intervention on the absolute level of student posttest performance would be similar across studies. Finally, we were interested in establishing whether the outcomes of students in the comparison groups, who received different SDI, would be comparable across studies on reading and reading-related measures.

Similar to findings reported in Simmons, Coyne, Hagan-Burke, Kwok, Simmons, Johnson et al. (2011), results from combined HLM models that included students from both the initial and varied replication studies indicated that during the initial study conducted in Texas and Connecticut, there was a statistically significant impact of ERI compared to the school-designed comparison condition on students' timed and untimed phonemic awareness outcomes, letter-sound knowledge, and untimed word attack outcomes, with effects sizes ranging between .40 and .51.

In the varied replication study conducted in Florida, however, no statistically significant differences between the ERI and SDI conditions were found on any posttest measure. In other words, the impact of ERI in the initial study did not replicate in the Florida study; that is, students in the treatment condition in the latter study did not benefit more than students in the comparison condition who received SDI. In fact, a trend favored the SDI condition on a number of the measures.

To unpack these findings, we next compared the absolute level of posttest performance of the students who received ERI during the initial and varied replication studies. Except for one statistically significant difference favoring the ERI students in the Florida replication on phonemic segmentation fluency ($g = .50$), there were no differences across studies. These findings suggest that the response of students to ERI was similar across the two studies, with students experiencing comparable achievement outcomes even though the ERI group outperformed the SDI comparison group in the initial study but not in the replication study.

Finally, we compared the outcomes of students who received the comparison SDI during the initial study with students who received SDI during the Florida varied replication study. In this case, analyses revealed a number of differences favoring the SDI students in the replication study. Specifically, we found statistically significant differences supporting Florida SDI students on both timed and untimed measures of phonemic awareness, letter-sound knowledge, nonsense word fluency, and word identification, with effect sizes ranging from .24 to 1.06. In addition, there was a trend favoring the Florida SDI students on all other measures as well. In other words, even though the comparison students in both studies had very similar pretest scores, the students who received SDI in the Florida replication study substantially outperformed the students who received SDI in the initial study.

In summary, results from combined analyses that included students from both the initial and varied replication studies suggest that differences in the impact of the ERI intervention across the initial and the replication study were largely explained by differences in students' response to the SDI in the two studies. In the initial study, compared to SDI, ERI had a substantively important impact on the majority of reading and reading-related measures. In the replication study, however, even though students who received ERI performed similarly to students who received ERI in the initial study, there was no differential impact of ERI because of the strong response of students to the SDI.

Differences in Instructional Context Across Studies

The stronger response of students to SDI in the varied replication study compared to the initial study may be explained in a number of ways, primarily related to the instructional

context in the Florida school district. First, kindergarten teachers in this district had received coordinated professional development over the previous years, which included summer reading academies and follow-up local workshops. This professional development focused on essential components of beginning reading instruction and evidence-based instructional strategies in early literacy. The school district had also been implementing common-core reading instruction using a published basal program, and teachers received support in delivering the instruction by school-based coaches. These efforts resulted in a coordinated focus on supporting beginning reading, a consistent approach to classroom instruction, and a common understanding of evidence-based practices in early literacy.

The Florida schools' coordinated approach to beginning reading instruction provided a strong foundation for delivering supplemental intervention to students who required additional support. In these schools, the provision of school-designed kindergarten intervention was part of accepted practice and implemented as a matter of course and policy (Just Read Florida, 2005). Moreover, because kindergarten teachers served as interventionists in the replication study, the SDI teachers were familiar and fluent with the SDI offered in the comparison condition. In contrast, ERI was new to the Florida interventionists.

Unlike the consistent framework for supporting beginning reading in the Florida replication schools, the Texas and Connecticut schools in the initial study took a more eclectic and individualized approach to reading instruction and intervention. Intervention schools were from three districts in Texas and three districts in Connecticut; therefore, the focused, district-level approach to early reading was not evident. Although some schools used published basal programs, implementation was less coordinated. Further, other schools used a combination of published materials and guided-reading strategies. In addition, professional development efforts in literacy varied among schools and across districts.

In the initial study, we sought to standardize a number of common instructional components across conditions by having comparable group sizes and asking interventionists to provide small-group instruction daily. However, for many of these schools, providing intervention in kindergarten was a novel idea. The structure of the ERI program may have provided support that facilitated implementation for interventionists in these schools. On the other hand, interventionists in the initial study had less experience than those in the replication study in developing and delivering the type of SDI that was provided to comparison students.

Program differentiation analyses also highlighted important instructional differences between the SDI provided to students in the initial study and the replication study. Although, in general, the content focus was similar in the SDI condition during the initial and the replication study, there were some clear and meaningful differences in instructional emphasis across the two studies. Compared to SDI in the initial study, the SDI instruction during the replication study included a greater emphasis on the phonological skills of blending and segmenting, sight word work, reading connected text, and writing sounds and words. These differences suggest that the SDI instruction provided to students during the replication study focused on more advanced reading and spelling skills.

IMPLICATIONS AND DIRECTIONS FOR FUTURE RESEARCH

We believe that the findings of our study have important implications for researchers designing and conducting intervention studies as well as those interpreting the results of such studies. First, our findings reinforce the importance of conducting varied replication studies of interventions in different settings and instructional contexts as well as the importance of

interpreting the results of single studies cautiously. We found a distinctly different pattern of results in the varied replication study than in the initial experiment. In the initial study, ERI had an overall impact on students compared to students receiving SDI. In our varied replication study, however, there was no differential impact of ERI compared to SDI with another group of schools in a different geographical location. This was the case even though students in both studies who received ERI performed similarly on outcome measures after adjusting for pretest scores and demographic differences.

Clearly, the findings of these two studies taken together require a much more careful and nuanced interpretation than the findings of the initial study in isolation. We believe that, in general, the results of multiple studies provide a more accurate, albeit more complex, picture of the overall effects of an intervention.

We found that the differences in the comparative impact of the ERI intervention across the initial and varied replication studies were largely explained by differences in students' response to the instruction provided to the comparison groups in the two studies. This finding underscores the importance of carefully considering what type of instruction comparison groups receive, because the nature and quality of the instruction provided to the control group is likely to influence relative treatment effects. Although no-treatment control groups provide a pure estimate of the impact of an intervention compared to nothing, for several reasons, this type of comparison is becoming less realistic and less feasible in beginning reading intervention research and less meaningful in interpreting intervention effects. First, "business as usual" instruction, which has often been interpreted as equivalent to a no-treatment control, in most schools now typically includes some type of supplemental intervention. Also, given the extensive and consistent evidence regarding the benefit of early reading intervention, it has become difficult to justify withholding intervention supports to control group participants.

We believe that it is also important to carefully and systematically describe the instruction provided to students in comparison groups. Such descriptions should go beyond a general overview sketched out in broad terms. Instead, multiple aspects of instruction should be observed and documented, including content focus, quality of instructional delivery, and dosage. With these data, it becomes possible to conduct program differentiation analyses (Dane & Schneider, 1998; Gersten et al., 2005) that allow for careful examination of actual differences across treatment and comparison conditions and support the isolation of critical components or "active ingredients" that account for intervention effects and impacts (Cordray & Pion, 2006).

Although we hypothesized that the differences in the impact of the intervention across studies were related to differences in the instructional context among the school districts in the efficacy and replication studies, we were unable to experimentally test this hypothesis. Other research teams have also suggested that contextual differences, particularly the quality and consistency of instruction among schools and districts may influence intervention effects (e.g., Lemons, 2010; McMaster et al., 2010). Therefore, future intervention research should consider designing varied replication or scale-up studies that systematically manipulate, or account for, important variables related to differences in instructional context across schools or research sites. Such studies could help isolate the contextual and instructional factors that are most highly associated with differences in impact as well as the extent to which interventions are robust to contextual variability. Moreover, they would provide educators with important information to support decisions related to adopting and implementing interventions by indicating whether an intervention is likely to positively influence the achievement of their students given the specifics of their school's instructional context and current practices. For example, the Florida district where the replication

study was conducted had invested significantly to ensure that many instructional factors that positively influence early reading success were solidly in place, including providing small-group kindergarten intervention designed around evidence-based practices and materials. In this case, investing extra resources in an additional, potentially costly, commercial program to replace the already-established intervention practices and materials may not affect student outcomes, at least in the short term.

However, in another district with fewer established beginning reading supports, adopting an explicit and systematic intervention may make a critical difference in improving reading achievement.

CONCLUSIONS

In this study, we found that the impact of supplemental beginning reading intervention observed in an initial randomized control study did not replicate in an experimental study conducted a year later in a different geographical location and a different instructional context. Multiyear analyses revealed that the differences in impact were largely explained by differences in students' response to the SDI that comprised the comparison condition in each study, rather than differential response to the experimental intervention. These findings suggest that factors related to instructional context may play an important role in the impact of beginning reading interventions and reinforce the value of varied replication studies for providing evidence about the effects of practices in different settings and under different conditions. Moreover, our findings suggest that determining the effects of interventions is more complex than straightforward.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324E060067 to Texas A&M University. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

REFERENCES

Educational Sciences Reform Act of 2002, Pub. L. 107–279, 116 Stat. 1939 (2002).

Beck, I. L., Farr, R. C., & Strickland, D. S. (2007). *Harcourt trophies*. Orlando, FL: Harcourt.

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*, *57*, 289–300.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). Washington, DC: American Psychological Association. doi:10.1037/11384-006

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.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test–Third Edition (PPVT-III)*. Bloomington, MN: Pearson Assessments.

Florida Center for Reading Research. (2005). *Kindergarten and first grade student center activities*. Tallahassee: Florida Department of Education.

Fritz, J. M., & Cleland, J. A. (2003). Effectiveness versus efficacy: More than a debate over language. *Journal of Orthopaedic and Sports Physical Therapy*, *3*, 163–165.

Fuchs, D., Fuchs, L. S., Thompson, A., Al Otaiba, S., Yen, L., McMaster, K., et al (2001). *Peer Assisted Learning Strategies: Kindergarten reading*. Nashville, TN: Vanderbilt University.

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

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills (6*th *ed.)*. Eugene, OR: Institute for the Development of Educational Achievement.

Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M. E., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice*, *15*, 198–205.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370. doi:10.3102/1076998606298043

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Just Read, Florida. (2005). K-12 comprehensive research-based reading plan state board rule. Retrieved from http://www.justreadflorida.com/docs/6A-6–053.pdf

Lemons, C. J., Fuchs, D., & Fuchs, L. S. (2008, February). *Evidence of kindergarten children's improved reading performance across 9 years in an urban school district: Implications for intervention research and educational policy*. Poster presented at the Pacific Coast Research Conference, Coronado, CA.

McMaster, K. L., Fuchs, D., Saenz, L., Lemons, C. J., Kearns, D., Yen, L., . . . Fuchs, L. S. (2010). Scaling up PALS: Importance of implementing evidence-based practice with fidelity and flexibility. *New Times for DLD: Division for Learning Disabilities*, *28*, 1–4.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development. Retrieved from http://www.nationalreadingpanel.org/

National Research Council. (2002). *Scientific research in education* (R. Shavelson & L. Towne, Eds.). Washington, DC: National Academy Press.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, *96*, 265–282. doi:10.1037/0022-0663.96.2.265

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for general causal inference*. Boston, MA: Houghton Mifflin.

Simmons, D. C., Coyne, M. D., Hagan-Burke, S., Kwok, O., Simmons, L., Johnson, C., . . . Crevecoeur, Y. (2011). Effects of supplemental reading interventions in anthertic contexts: A comparison of kindergartners' response. *Exceptional Children*, *77*, 207–228.

Van IJzendoorn, M. H. (1994). *A process model of replication studies: On the relation between different types of replication*. Leiden, the Netherlands: Leiden University Library.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed.

What Works Clearinghouse. (2008). *Procedures and standards handbook, version 2.0* [Electronic version]. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf

Woodcock, R. W. (1987, 1998). *Woodcock Reading Mastery Tests–Revised/Normative update*. Bloomington, MN: Pearson Assessments.