


The Relationship Between Study Quality and the Effects of Supplemental Reading Interventions: A Meta-Analysis

Exceptional Children
2019, Vol. 85(3) 347–366
© The Author(s) 2018
DOI: 10.1177/0014402918796164
journals.sagepub.com/home/ecx


Christy R. Austin¹, Jeanne Wanzek², Nancy K. Scammacca¹, Sharon Vaughn¹, Samantha A. Gesel², Rachel E. Donegan², and Morgan L. Engelmann¹

Abstract

Empirical studies investigating supplemental reading interventions for students with or at risk for reading disabilities in the early elementary grades have demonstrated a range of effect sizes. Identifying the findings from high-quality research can provide greater certainty of findings related to the effectiveness of supplemental reading interventions. This meta-analysis investigated how four variables of study quality (study design, statistical treatment, Type I error, and fidelity of implementation) were related to effect sizes from standardized measures of foundational reading skills and language and comprehension. The results from 88 studies indicated that year of publication was a significant predictor of effect sizes for both standardized measures of foundational reading skills and language and comprehension, with more recent studies demonstrating smaller effect sizes. Results also demonstrated that with the exception of research design predicting effect sizes on foundational reading skills measures, study quality was not related to the effects of supplemental reading interventions. Implications for research and practice are discussed.

Keywords

study quality, reading intervention, meta-analysis

Evidence suggests that educational reform has resulted in students with learning disabilities receiving inclusion support, with a focus on universal design for learning and accommodations to give students access to high standards in the general education classroom (Fuchs et al., 2015). Despite inclusion reform, data from the National Assessment of Education Progress (National Center for Education Statistics, 2011, 2013, 2015) demonstrates a large achievement gap between students with disabilities and their nondisabled peers that continues over time. The percentages of students with disabilities scoring at or above the basic level in reading in 2011, 2013, and 2015

were 32%, 31%, and 33%, respectively. For students without disabilities, these numbers were 67%, 68%, and 69%. The magnitude and severity of this achievement gap suggests that students with disabilities might benefit from more intensive support than what is currently

¹University of Texas at Austin

²Vanderbilt University

Corresponding Author:

Christy R. Austin, Meadows Center for Preventing Educational Risk, College of Education SZB 228, University of Texas at Austin, 1912 Speedway D4900, Austin, TX 78712-1284.

E-mail: Christyaustin@utexas.edu

provided in order to better assist them with accessing the high standards in the general education classroom.

Supplemental reading interventions, or instruction provided in addition to core instruction in the general education classroom, can remediate student reading difficulties and prevent school failure (Mathes et al., 2005; O'Connor, Bocian, Beach, Sanchez, & Flynn, 2013; O'Connor Fulmer, Harty, & Bell, 2005; Vaughn et al., 2009; Wanzek et al., 2016). In addition, supplemental interventions can help to identify students with more significant difficulties and disabilities who are in need of more intensive support (Carney & Stiefel, 2008; O'Connor et al., 2013; Wanzek et al., 2016). In order to effectively implement supplemental reading interventions, it is important to be able to identify evidence-based practices (EBPs) that can increase the likelihood that students will respond adequately to instruction.

Identifying EBPs

An EBP can be defined as an instructional strategy, intervention, or teaching program that has resulted in consistent positive results when tested in scientific research (Council for Exceptional Children, 2015). Professional standards (Council for Exceptional Children, 2015) and U.S. federal regulations of the Every Student Succeeds Act (2015–2016) reauthorized by the Elementary and Secondary Education Act mandate the use of EBPs for students with disabilities. In order to identify a practice as evidence based, it is necessary to investigate the quality of scientific research.

However, identifying the quality of a study is complicated by the fact that there are a variety of entities responsible for evaluating programs and interventions as evidence based. Government agencies (e.g., U.S. Department of Education's What Works Clearinghouse), professional organizations (e.g., Council for Exceptional Children's Research Division, School Psychology Division of the American Psychological Association), information clearinghouses (e.g., Cochrane Group, the National Academies, Scottish Intercollegiate Guidelines Network), and other groups have

their own criteria to evaluate the quality of a study, and the process or weighting of criteria can vary across groups (Goldstein, Lackey, & Schneider, 2014). When synthesizing across studies and interpreting findings, considering study quality can allow for more accurate identification of EBPs in a particular domain, as there is increased clarity regarding the findings consistently demonstrated by high-quality research.

Current Evidence on the Effects of Intensive Early Reading Interventions

Much of the research on early elementary reading interventions has been synthesized to provide educators with summaries of the effects of various implementations and intensities of reading interventions for students with or at risk for reading disabilities, or a specific learning disability in the area of reading as defined by the Individuals With Disabilities Education Act. Extensive evidence demonstrates the benefit of reading instruction at the early elementary level targeting both foundational reading skills, such as phonological awareness, phonics, word recognition, and reading fluency, as well as higher-order skills, such as language, vocabulary, and comprehension (National Early Literacy Panel, 2008; National Reading Panel, 2000). Specifically, students with and at risk for reading disabilities benefit from direct instruction in foundational reading skills and from including strategy instruction with direct instruction in reading comprehension (Swanson & Hoskyn, 2000; Swanson, Hoskyn, & Lee, 1999; Torgesen et al., 2017). Finally, prior research has demonstrated that pretest scores moderate student responsiveness to supplemental reading interventions at the elementary level (Tran, Sanchez, Arellano, & Swanson, 2011).

Wanzek and Vaughn (2007) synthesized extant research published between 1995 and 2005 investigating supplemental reading interventions provided for 100 or more sessions for students with reading difficulties and disabilities in Grades K–3. Studywise mean effects ranged from -0.05 to 0.84 and were summarized by examining effects for the

duration of intervention, instructional group size, grade level, and type of intervention. Findings indicated that few differences were seen in the magnitude of effect sizes based on the duration of instruction. Studies implemented one-on-one were generally associated with higher effects (mean effect size range 0.17 to 0.84) than studies implemented in small groups of two to eight students (mean effect size range -0.05 to 0.39). In addition, interventions provided in Grades K–1 were associated with higher effects than those provided in Grades 2–3.

Wanzek et al. (2016) conducted a meta-analysis examining the effects of reading interventions implemented for 15 to 99 sessions on the foundational reading skills, language, and comprehension of students with or at risk for reading difficulties in Grades K–3. Overall, the results demonstrated moderate, positive effect sizes of supplemental reading interventions for struggling readers in Grades K–3. The mean effect size was 0.54 on standardized measures of foundational skills and 0.62 on researcher-developed measures of foundational skills. On measures of language and comprehension, researchers reported a mean effect size of 0.36 on standardized measures and a mean effect size of 1.02 on researcher-developed measures. Results also indicated that there were no differences in the magnitude of effects based on differences in intervention type, instructional group size, grade level, intervention implementer, or the number of hours of intervention.

Wanzek et al. (2018) updated and extended the synthesis conducted by Wanzek and Vaughn (2007). The previous work was updated by identifying studies published after 2005 and was extended by conducting a meta-analysis with the larger corpus of studies to analyze the effects of supplemental interventions provided for 100 or more sessions. Results demonstrated that the effects of early reading interventions for students with or at risk for reading disabilities were similar to the results from the meta-analysis of early reading interventions implemented for 15 to 99 sessions. Findings indicated significant, positive effects, with a weighted mean effect size of 0.39 and a mean effect size of 0.28 after

adjusting for publication bias. There were no significant differences in the magnitude of effects based on differences in intervention type, instructional group size, grade level, implementer, or total hours of intervention.

Purpose and Research Questions

The findings from each of these syntheses and meta-analyses support the early implementation of supplemental interventions for students with or at risk for reading disabilities. In order to extend the previous research investigating supplemental interventions for early struggling readers, the aim of the current study is to reevaluate the results of the prior meta-analyses to determine how study quality relates to student outcomes. In order to evaluate the relationship between study quality and effect sizes, we also investigated the relationship between year of publication and effect sizes, given the knowledge that implementation guidelines and study designs have evolved and become more rigorous over time. This meta-analysis includes studies from the Wanzek et al. (2016) and Wanzek et al. (2018) meta-analyses as well as studies published after these meta-analyses were completed.

We used a set of quality indicators adapted from a rubric used in a systematic review conducted by Goldstein et al. (2014). Goldstein et al. (2014) analyzed what effect study quality had on the social-skill outcomes of preschoolers with autism spectrum disorder (ASD). A wide range of effect sizes had been demonstrated previously, convoluting the conclusions that could be drawn about identifying EBPs for improving social-skill outcomes for this population of students. By utilizing a new approach for evaluating study quality, Goldstein et al. demonstrated that although the overall quality of studies fell short both within and across studies, there was sufficient research to determine that social skills interventions held promise as an EBP for preschoolers with ASD. This work demonstrates the need for use of a similar approach for evaluating research topics that have generated a wide range of effect sizes, making it difficult to draw conclusions regarding best practices for improving student outcomes.

To increase reliability in scoring, we revised the approach for evaluating study quality utilized in Goldstein et al. (2014) by reducing the quality ratings to three categories (exemplary, acceptable, unacceptable) rather than four (exemplary, acceptable, minimal, unacceptable). In general, we combined the *acceptable* and *minimal* categories into one category of *acceptable*. We evaluated each study in the areas of (a) study design, (b) implementation fidelity, (c) statistical analysis, and (d) likelihood of Type I error. The purpose of this meta-analysis is to synthesize recent reading interventions for students with or at risk for reading disabilities, analyzing the association of the quality of intervention studies with the magnitude of treatment effects on reading outcomes for students with reading difficulties in Grades K–3. Specifically, the research questions investigated were as follows: (1) What is the quality of early reading intervention studies completed in the past 20 years? (2) How does the quality of K–3 reading intervention research for students with reading difficulties and disabilities relate to student outcomes?

The purpose of this meta-analysis is to synthesize recent reading interventions for students with or at risk for reading disabilities, analyzing the association of the quality of intervention studies with the magnitude of treatment effects on reading outcomes for students with reading difficulties in Grades K–3.

Method

In order to identify the studies included in this meta-analysis, we followed a two-step search procedure. First, we included studies from the previous meta-analyses (Wanzek et al., 2016, 2018) that met all inclusion criteria. A total of 80 articles (82 studies) were included. In addition, we updated the searches of ERIC and PsycINFO using the same search terms to identify studies completed in reading inter-

ventions (*reading interven**, *reading instruction*, *reading strategies*, *supplemental instruction*, *special educ**, *phon**, *fluency*, *vocab**, *comp**) with our population of interest (*reading difficult**, *learning disab**, *reading disab**, *reading delays*, *reading disorder**, *dyslex**) to result in a corpus of studies meeting criteria from the years 1995 to 2016. The updated search yielded 4,342 abstracts. As with the previous searches, our keywords identified many abstracts from research in other disciplines (e.g., aphasia, dementia) that are related to terms such as *delays*, *disability*, *fluency*, and *comprehension*. To participate in the screening and sorting process, coders were required to reach 100% reliability in decisions regarding abstracts for the first 200 abstracts before continuing their search. All individuals reached 100% accuracy. Abstract information eliminated 4,241 articles. We examined the full text of the remaining articles ($n = 101$) and found an additional six articles that met all selection criteria for the meta-analysis. This yielded a total of 86 articles detailing 88 studies that met all criteria and were included in this review. See Figure S1 for a flow chart of search and inclusion results. We included articles that met the following criteria:

1. The article was written in English and published in a peer-reviewed journal.
2. Participants were students in Grades K–3 identified with a learning disability, with reading difficulty, or at risk (e.g., students with low reading fluency, deficits in phonological awareness, below-average reading or language achievement). Studies with additional participants were included if more than 50% of the participants were part of the targeted population or if disaggregated data were provided for participants in the targeted population.
3. Interventions targeted literacy in an alphabetic language and were provided in a school setting. Home and clinic interventions were not included.
4. Intervention was provided for a minimum of 15 sessions and was not part

of the general education curriculum provided to all students.

5. At least one standardized reading measure of phonological awareness, phonics, word recognition, fluency, vocabulary, oral language, or reading comprehension was used.
6. The study used an experimental or quasiexperimental design, and data were provided to calculate effect sizes.

Coding Procedures

To organize information about each study, we adapted a coding sheet based on quality indicators for systematic research reviews as presented by Goldstein et al. (2014). We specifically examined study design, statistical treatment, Type I error, and fidelity of implementation. For each area, we used key indicators to assign a rating of *unacceptable*, *acceptable*, or *exemplary* on a 3-point Likert-type scale (0 = *unacceptable*, 1 = *acceptable*, 2 = *exemplary*). The quality areas and indicators used to assign ratings are presented in Table 1.

Most of the included studies ($n = 59$) had a single contrast of a treatment and comparison condition. The remaining 29 studies reported data for multiple treatment or comparison conditions. We coded the quality of each contrast within a study, for a total of 127 contrasts. For example, if a study had two treatment groups and one comparison group, we coded the quality of the first treatment group versus comparison contrast and the second treatment group versus comparison contrast. Table 2 reports the quality coding scores for every contrast from the included studies across the quality criteria. For studies that were also included in the updated meta-analyses, we collected information from code sheets completed during the previous meta-analyses to calculate effect sizes.

Four trained graduate students completed all coding. The second author trained coders on all indicators and rating criteria and served as the gold standard in ensuring that all coders were reliable prior to coding the included studies (Gwet, 2001). Coders demonstrated a minimum of 90% initial reliability to the gold

standard before beginning coding. Mean reliability among the four coders was 97.7%. Two coders independently coded each study. The two coders resolved any discrepancies in ratings by reviewing rating criteria for the indicator in question and discussing until reaching a consensus.

Effect Size Calculation

We computed Hedges' g effect sizes using the means and standard deviations for the treatment and comparison groups in studies where these data were reported. Studies that lacked this information reported Cohen's d effect sizes and sample sizes for the treatment and comparison groups, allowing for calculating Hedges' g effect sizes based on these statistics. We used the Comprehensive Meta Analysis (Version 3.3.070) software (Borenstein, Hedges, Higgins, & Rothstein, 2013) to compute all effect sizes.

Meta-Analysis Procedures

The meta-analysis included only those studies that reported outcomes on standardized measures ($k = 88$) in order to control for differences due to measurement quality and based on findings from previous intervention research indicating that effect sizes from standardized and unstandardized measures differed (Scammacca, Roberts, Vaughn, & Stuebing, 2015; Swanson et al., 1999; Willingham, 2007). We did not analyze supplemental interventions implemented for 15 to 99 sessions separately from interventions implemented for 100 or more sessions due to the previous meta-analyses that demonstrated similar findings regardless of the number of intervention sessions. To limit heterogeneity due to the reading domain measured, we conducted separate meta-analyses on effect sizes from measures of foundational reading skills (including phonological awareness, decoding, word identification, decoding fluency, word identification fluency, text reading fluency, and spelling) and measures of language and comprehension (including vocabulary, oral language, listening comprehension, and reading comprehension). Of the 88 studies

Table 1. Quality Indicators and Rating Criteria.

Quality area	Rating	Rating criteria
Design	Exemplary	A randomized design with a sufficiently large sample (≥ 20) from a clearly described population.
	Acceptable	A randomized design with insufficient sample size (< 20) or a nonrandomized design (quasiexperimental study) with a large sample and evidence the groups are equivalent prior to the study based on pretest scores.
	Unacceptable	Nonrandomized design with small sample or lack of evidence that groups equivalent prior to the start of the study based on pretest scores.
Implementation fidelity	Exemplary	Clear, replicable operational definitions of treatment procedures, high procedural fidelity ($\geq 75\%$), and interobserver reliability data ($\geq .90$).
	Acceptable	Clear, replicable operational definitions of treatment procedures, high procedural fidelity ($\geq 75\%$), and interobserver reliability data ($\geq .80$).
	Unacceptable	Unreplicable description of treatment, poor implementation fidelity ($< 75\%$), poor intercoder agreement ($< .80$), or fidelity was not reported.
Statistical analyses	Exemplary	Appropriate use of analysis matching the design of the study; the unit of assignment to condition (e.g., student, class, school) matches the unit of analysis or clustering or nesting of students in the unit of assignment is taken into account at analysis; sufficiently large sample size (≥ 20), and effect sizes are reported.
	Acceptable	Appropriate use of analysis matching the design of the study; the unit of assignment to condition (e.g., student, class, school) matches the unit of analysis or clustering or nesting of students in the unit of assignment is taken into account at analysis; sufficiently large sample size (≥ 20), adequate information to determine effect size if not reported.
	Unacceptable	Inappropriate use of analysis or insufficient sample size (< 20).
Likelihood of Type I error	Exemplary	When multiple comparisons are conducted, the p value is adjusted to control Type I error.
	Acceptable	Low likelihood of Type I error based on lack of significance in tests or Benjamini-Hochberg correction results.
	Unacceptable	High likelihood of Type I error based on Benjamini-Hochberg correction results or inadequate data to apply Benjamini-Hochberg correction.

included in this synthesis, 85 included standardized measures of foundational skills and 49 included standardized measures of language and comprehension.

In the meta-analysis of standardized measures of foundational reading skills, 78 of the 85 studies contributed multiple effect sizes, resulting in a total of 516 effect sizes available for the analysis. In the meta-analysis of standardized measures of language and comprehension, 31 of 49 studies contributed multiple effect sizes, resulting in a total of 96 effect sizes for analysis. The presence of multiple effect sizes in a study usually resulted from

multiple measures being used to determine the treatment effect. However, multiple effect sizes reported in some studies also resulted from the inclusion of more than one pair of treatment-comparison group contrasts or multiple subgroup comparisons (e.g., when results were reported by grade for multiple grades). Multiple effect sizes within a study from any of these three sources are dependent, meaning that they are correlated to some degree; the meta-analysis therefore must account for this dependence to provide unbiased estimates of the mean effect size and its standard error.

Table 2. Quality Coding Scores Across Indicators.

Author (year)	Average rating by contrast	Design	Implementation fidelity	Statistical analyses	Likelihood of Type I error
Al Otaiba et al. (2005)	1.25	○	⊙	○	●
Baker et al. (2000)	1.25	○	●	○	⊙
Barker & Torgesen (1995)					
Daisy Quest vs. BAU	1.25	○	●	○	⊙
Hint and Hunt vs. BAU	1.00	○	●	●	○
Berninger et al. (2006)	0.50	●	⊙	⊙	●
Berninger et al. (2003)					
Word recognition (rec.) training vs. reading practice	0.75	○	●	●	⊙
Word rec. training vs. BAU	0.75	○	●	●	⊙
Comprehension (comp.) training vs. reading practice	1.00	○	●	○	●
Comp. training vs. BAU	0.50	○	●	●	●
Word rec. and comp. training vs. reading practice	0.25	⊙	●	●	●
Word rec. and comp. training vs. BAU	1.50	○	●	○	○
Brown et al. (2005)	1.00	●	⊙	○	⊙
Burns et al. (2004)	0.75	●	⊙	○	●
Catts et al. (2015)	0.25	●	●	⊙	●
Center et al. (1995)	0.75	●	⊙	●	○
Chapman et al. (2001)					
Repeated reading vs. comparison	0.50	●	⊙	⊙	●
Referred on vs. comparison	1.50	○	○	⊙	⊙
Coyne et al. (2013)	0.75	●	⊙	⊙	⊙
Denton et al. (2014)					
Guided reading vs. BAU	0.25	●	⊙	●	●
Explicit instruction vs. BAU	1.50	○	○	○	●
Denton, Nimon, et al. (2010)	1.50	○	○	○	●
Denton, Solari, et al. (2010)	1.75	○	○	○	⊙
Duff et al. (2014)	1.00	○	⊙	⊙	●
Ehri et al. (2007)					
Reading Rescue tutoring vs. Voyager intervention	1.50	○	⊙	⊙	○
Reading Rescue tutoring vs. class reading	0.75	⊙	●	⊙	⊙
Fawcett et al. (2001)	1.00	○	⊙	●	⊙
Fien et al. (2015)	1.50	○	⊙	○	⊙
Foy (2009)	1.75	○	⊙	○	○
Fuchs & Fuchs (2006)	0.00	●	●	●	●
Gilbert et al. (2013)	0.75	⊙	○	●	●
Gillon (2000)					
Phonological awareness vs. traditional	0.50	○	●	●	●
Phonological awareness vs. minimal	0.00	●	●	●	●
Graham et al. (2002)	0.00	●	●	●	●
Gunn et al. (2000)	0.75	⊙	⊙	●	⊙

(continued)

Table 2. (continued)

Author (year)	Average rating by contrast	Design	Implementation fidelity	Statistical analyses	Likelihood of Type I error
Hagan-Burke et al. (2011)	0.75	⊙	⊙	●	⊙
Hatcher et al. (2006)	1.50	○	⊙	○	⊙
Hurry & Sylva (2007)					
Phonological training vs. BAU	1.50	○	⊙	⊙	○
Phonological training vs. between school BAU	1.00	⊙	●	⊙	○
Jenkins et al. (2004)					
More decodable texts vs. BAU	0.75	⊙	●	●	○
Less decodable texts vs. BAU	0.75	⊙	●	○	●
Kerins et al. (2010)	1.50	⊙	⊙	○	○
Kyle et al. (2013)					
GraphoGame phoneme vs. not	0.50	●	●	○	●
GraphoGame rime vs. not	0.75	●	●	○	⊙
Lane (1999)	1.00	○	●	○	●
Lane et al. (2009)					
UFLI vs. BAU	0.75	⊙	●	●	○
UFLI, no word work vs. BAU	0.75	●	●	⊙	○
UFLI, no sentence writing vs. BAU	0.75	●	●	⊙	○
UFLI, no literacy extension vs. BAU	1.50	○	⊙	⊙	○
Lee et al. (2011)	1.75	○	⊙	○	○
Lee & Scanlon (2015)	1.25	○	⊙	⊙	⊙
Lennon & Slesinski (1999)					
Tutoring-low vs. comparison-low	1.00	⊙	●	○	⊙
Tutoring-mid vs. comparison-mid	1.50	○	⊙	○	⊙
Little et al. (2012)	0.75	⊙	⊙	●	⊙
Marston et al. (1995)					
DI for SRA vs. nonequivalent comparison	1.50	⊙	⊙	○	○
CAI vs. nonequivalent comparison	1.25	⊙	⊙	⊙	○
Effective teaching vs. nonequivalent comparison	0.75	⊙	●	●	○
Mathes & Babyak (2001)	0.50	⊙	●	⊙	●
Mathes et al. (2005)					
Proactive vs. enhanced class	1.00	○	⊙	⊙	●
Responsive vs. enhanced class	1.00	⊙	●	○	⊙
Mathes et al. (2003)					
Peer Assisted Learning Strategies (PALS) vs. no treatment	0.50	●	●	⊙	⊙
Teacher-directed instruction vs. no treatment	0.75	⊙	●	⊙	⊙
McCarthy et al. (1995)	0.25	⊙	●	●	●
McMaster et al. (2005)					
Tutoring vs. PALS	0.00	●	●	●	●
Tutoring vs. Modified PALS	1.25	○	⊙	○	●
Meier & Invernizzi (2001)	0.75	○	●	●	⊙
Miller (2003)					
Partner in reading vs. comparison	0.75	○	⊙	●	●
Reading Recovery vs. comparison	0.50	○	●	●	●

(continued)

Table 2. (continued)

Author (year)	Average rating by contrast	Design	Implementation fidelity	Statistical analyses	Likelihood of Type I error
Mokhtari et al. (2015)	0.50	⊙	●	●	⊙
Morris et al. (2012)					
PHAST vs. Math + CSS	1.75	○	○	⊙	○
PHAB + RAVE-O vs. Math + CSS	1.25	○	○	⊙	●
PHAB + CSS vs. Math + CSS	1.25	○	⊙	○	●
Morris et al. (2000)	0.50	●	●	⊙	⊙
Nelson et al. (2005)	1.25	○	●	○	⊙
Nicolson et al. (1999)	1.50	⊙	○	○	⊙
Nielsen & Friesen (2012)	0.75	⊙	○	●	●
Osborn et al. (2007)					
Project MORE (SLD) vs. comparison SLD	0.75	○	⊙	●	●
Project MORE (Title I) vs. comparison Title I	1.00	○	●	⊙	⊙
O'Shaughnessy & Swanson (2000)					
Phonological awareness training vs. math	0.75	○	⊙	●	●
Word analogy training vs. math	0.75	○	⊙	●	●
Papadopoulos et al. (2003)	1.00	○	●	○	●
Pericola-Case et al. (2010)	0.75	○	⊙	●	●
Pullen & Lane (2014)					
Treatment with word work vs. BAU	0.75	○	●	⊙	●
Treatment without word work vs. BAU	1.00	○	⊙	⊙	●
Rashotte et al. (2001)	0.00	●	●	●	●
Reutzel et al. (2012)	1.00	⊙	⊙	○	●
Rimm-Kaufman et al. (1999)	0.00	●	●	●	●
Ryder et al. (2008)	1.00	●	⊙	⊙	○
Santa & Høien (1999)	1.25	⊙	○	⊙	⊙
Schwartz (2005)	1.25	○	⊙	●	○
Simmons et al. (2011)	0.50	●	⊙	●	⊙
Torgesen et al. (1997)					
Regular classroom reading vs. no treatment	0.75	○	●	●	⊙
PASP vs. no treatment	0.25	●	●	⊙	●
Embedded phonics vs. no treatment	0.25	●	●	⊙	●
Torgesen et al. (2010)					
Read, Write, and Type vs. BAU	0.25	●	⊙	●	●
Lindamood Phoneme Sequencing Program vs. BAU	0.25	⊙	●	●	●
Vadasy et al. (1997a)	1.00	⊙	⊙	⊙	⊙
Vadasy et al. (1997b)	0.50	○	●	●	●
Vadasy et al. (2000)	0.50	○	●	●	●
Vadasy & Sanders (2008a)	0.75	⊙	●	⊙	⊙
Vadasy & Sanders (2008b)					

(continued)

Table 2. (continued)

Author (year)	Average rating by contrast	Design	Implementation fidelity	Statistical analyses	Likelihood of Type I error
Individual tutoring vs. no tutoring	0.00	●	●	●	●
Dyad tutoring vs. no tutoring	1.00	⊙	○	●	⊙
Vadasy & Sanders (2009)					
Teacher treatment vs. comparison	1.50	○	○	○	●
Paraprofessional treatment vs. comparison	1.50	○	⊙	○	⊙
Vadasy & Sanders (2010)	1.25	○	⊙	⊙	⊙
Vadasy & Sanders (2011)	1.25	⊙	●	○	○
Vadasy et al. (2005)					
Reading practice vs. comparison	0.50	⊙	●	●	⊙
Word study vs. comparison	0.25	⊙	●	●	●
Vadasy et al. (2006a)	1.00	●	●	○	○
Vadasy et al. (2006b)					
Study 1	0.50	●	●	○	●
Study 2	1.00	⊙	⊙	○	●
Vadasy et al. (2002)					
Sound Partners + Thinking Partners vs. BAU	0.75	○	●	●	⊙
Thinking Partners vs. BAU	1.00	⊙	○	●	⊙
Vadasy et al. (2007)	1.00	⊙	⊙	○	●
Vaughn et al. (2006)	1.00	○	●	○	●
Vellutino et al. (2008)	0.75	⊙	●	○	●
Vernon-Feagans et al. (2012)					
Targeted reading intervention vs. comparison (Grade: K)	1.00	○	●	○	●
Targeted reading intervention vs. comparison (Grade: 1)	0.75	○	⊙	●	●
Wang & Algozzine (2008)	1.25	⊙	●	○	○
Wanzek & Vaughn (2008)					
Study 1	1.25	⊙	●	○	○
Study 2	0.50	⊙	●	⊙	●
Wise et al. (1999)					
Combination vs. comparison	0.75	⊙	●	○	●
Sound manipulation vs. comparison	0.75	⊙	●	○	●
Articulation vs. comparison	0.75	⊙	●	○	●
Wise et al. (2016)	0.75	⊙	●	○	●
Wright & Jacobs (2003)					
Phonological awareness training vs. comparison	1.00	⊙	●	○	⊙
Phonological awareness + MCMS vs. comparison	0.50	○	●	●	●
Zvoch, & Stevens (2013)	1.00	○	⊙	●	⊙
Average score by indicator	0.87	1.22	0.57	0.98	0.70

Note. See online supplementary materials for references included in the meta-analysis. 2 = ○ Exemplary; 1 = ⊙ Acceptable; 0 = ● Unacceptable. BAU = business as usual; CAI = computer-assisted instruction; CSS = classroom survival skills; DI = direct instruction; MCMS = metalinguistic concepts and metacognitive strategies; MORE = Mentoring in Ohio for Reading Excellence; PALS = Peer Assisted Learning Strategies; PASP = phonological awareness and synthetic phonics; PAT = phonological awareness training; PHAB = phonological analysis and blending; PHAST = Phonological and Strategy Training; RAVE-O = Retrieval, Automaticity, Vocabulary, Engagement With Language, and Orthography; SLD = specific learning disability; SRA = Science Research Associates; UFLI = University of Florida Literacy Initiative.

To accommodate the dependency in the meta-analytic data sets, we implemented robust variance estimation (RVE; Hedges, Tipton, & Johnson, 2010) to estimate meta-regression models using the *robumeta* package for R (Fisher & Tipton, 2015) to calculate beta coefficients, mean effect sizes, and standard errors. We conducted hypothesis tests for categorical moderators using the *clubSandwich* package for R (Pustejovsky, 2015). Based on recommendations from Tipton and Pustejovsky (2015), we implemented the small-sample correction in all models to avoid inflating Type I error (Tipton, 2015). To implement RVE, the mean within-study correlation between all pairs of effect sizes (ρ) must be specified to allow for estimation of appropriate study weights and to calculate between-study variance. However, as shown by Hedges et al. (2010), the value used for ρ does not alter the results meaningfully; they recommended conducting a sensitivity analysis testing the impact of different ρ values on the model parameters. Therefore, we tested .2, .5, and .8 as values for ρ . No meaningful differences were found in the results across models for either set of measures. The results reported below are from the models where $\rho = .8$.

We estimated meta-regression models for the meta-analyses of the standardized foundational reading skills and language and comprehension measures. First, we estimated a mixed-effects meta-regression model that included coefficients for four categorical moderators (quality of design, implementation fidelity, statistical treatment, and handling of Type I error) and one continuous moderator (year of study publication). We conducted tests of statistical significance for the difference between coefficients for the categorical moderators using Wald tests as recommended in Tipton and Pustejovsky (2015) and implemented in the *clubSandwich* package for R (Pustejovsky, 2015). This approach avoids inflating Type I error, especially when the number of studies at each level of the moderator is below 40. However, power for the moderator analyses is diminished as a result. Therefore, we chose to implement a $p < .10$ criterion for determining the statistical significance of model parameters. Finally, we

estimated intercept-only models to determine the weighted mean effect size and standard error for studies at each level of each moderator in each meta-analytic data set.

Analysis of Publication Bias

Because we did not include unpublished studies in this meta-analysis, publication bias is a potential threat to the validity of our results. To evaluate the potential impact of publication bias, we implemented the trim-and-fill method (Duval & Tweedie, 2000) using a fixed-effects model. The trim-and-fill method deletes effect sizes that produce asymmetry in the funnel plot of effect sizes, calculates a mean effect, and (if needed) imputes a sufficient number of effect sizes to create a symmetrical funnel plot. It then produces an effect size estimate that includes these missing studies. In this meta-analysis, publication bias was most likely to affect the number of studies with low quality that were available for inclusion, given that low-quality studies (especially those with small effects) are less likely to be published. To investigate whether studies were likely to be missing from among the set of studies with quality ratings of *unacceptable*, we analyzed publication bias by implementing the trim-and-fill method with the effect sizes from studies with unacceptable ratings.

Results

Quality of Early Reading Intervention Studies

Overall, the quality of early reading intervention studies for students with or at risk for reading disabilities completed in the past 20 years varied by study, quality area, and contrast within each study (see Table 2). As a set, the coded contrasts had a mean quality score of 0.87 (range: 0–1.75; unacceptable). The average score across contrasts by quality area was 0.57 for implementation fidelity (unacceptable), 0.70 for likelihood of Type I error (unacceptable), 0.98 for statistical analysis (nearing acceptable), and 1.22 for design (acceptable).

Table 3 provides an overview of the number of contrasts receiving each quality score.

Table 3. Quality Coding Across Studies.

Variable	Design	Implementation fidelity	Statistical analyses	Likelihood of Type I error
Number of contrasts receiving exemplary score	56 (44.1%)	13 (10.2%)	46 (36.2%)	24 (18.9%)
Number of contrasts receiving acceptable score	43 (33.9%)	46 (36.2%)	33 (26.0%)	41 (32.3%)
Number of contrasts receiving unacceptable score	28 (22.0%)	68 (53.5%)	48 (37.8%)	62 (48.8%)

Note. These values based on 127 contrasts from 88 studies published in 86 publications.

As a set, the greatest number of contrasts ($k = 56$; 44.1%) used randomized designs with large samples (i.e., 20 or more participants in each condition); an additional 43 contrasts (33.9%) used either a randomized design with a small sample size or a nonrandomized design but with good evidence of group equivalence. Only 28 of the contrasts (22.0%) had unacceptable design quality, indicating a lack of randomized design, with either insufficient evidence of group equivalence or a small sample size.

For the statistical analysis quality area, the majority of the contrasts used an appropriate statistical analysis ($k = 79$; 62.2%) to match the design of the study; 46 of these contrasts (36.2% of all contrasts; 58.2% of the contrasts with appropriate analysis) also reported effect sizes and were rated exemplary. The 33 contrasts rated acceptable for this indicator (26.0% of all contrasts; 41.8% of the contrasts with appropriate analysis) did not report effect sizes. Contrasts earned an unacceptable rating for the statistical analysis for one of two reasons: either the authors did not use an appropriate statistical analysis (e.g., did not account for nesting if randomization occurred at a level other than the student; $k = 21$; 16.5%) or sample size included fewer than 20 participants in each condition ($k = 27$; 21.3%).

Only 24 of the contrasts (18.9%) received an exemplary quality rating for likelihood of Type I error. To receive an exemplary quality rating in this area, either the study had only one dependent measure, eliminating the need to control for multiple comparisons, or the authors properly accounted for multiple comparisons in their analyses. If the authors did

not control for multiple comparisons in their analyses, we applied the Benjamini-Hochberg correction to determine the likelihood of Type I error. The majority of contrasts ($k = 62$; 48.8%) scored unacceptable because there was no control for multiple comparisons and the likelihood of Type I error was high after applying the Benjamini-Hochberg corrections. The remaining 41 contrasts (32.3%) were rated as acceptable because post hoc corrections showed a low likelihood of Type I error.

Finally, the majority of contrasts received an unacceptable score for the implementation fidelity quality area ($k = 68$; 53.5%). Contrasts scored unacceptable on this indicator for one of three reasons: (a) low fidelity scores reported ($k = 3$; 2.4%), (b) no reported fidelity information ($k = 31$; 24.4%), or (c) fidelity mentioned, but authors did not provide actual data required to determine the quality of implementation fidelity ($k = 34$; 26.8%). Only 16 contrasts (12.6%) scored exemplary on this indicator. The remaining 43 contrasts (33.9%) scored acceptable on this indicator. Most of these contrasts demonstrated high implementation fidelity but did not report interobserver reliability for the fidelity measure used.

Analyzing the quality of intervention studies was challenging due to the variation in quality ratings within studies and across contrasts. Eighty-five out of 88 studies (97%) received different quality ratings across the four quality indicators. In addition, 29 studies included multiple contrasts. Twenty-seven of these 29 (93%) studies involving multiple contrasts included different quality ratings between contrasts on at least one of the quality indicators.

Meta-Analysis of Effects From Measures of Foundational Reading Skills

The meta-regression of the standardized measures of foundational reading skills included 516 effect sizes from 85 studies. The model included the four quality variables as categorical moderators and year of study publication as a continuous moderator. Results indicated that studies with exemplary design quality were associated with smaller effect sizes ($b = -0.28$, $SE = 0.12$, $p = .035$). Year of publication also was a statistically significant predictor of smaller effect sizes ($b = -0.02$, $SE = 0.01$, $p = .004$). The I^2 estimate of the percentage of between-study heterogeneity not due to chance variation in effects was 68.88%, with a τ^2 estimate of the true variance in the population of effects of .11. See Table 4 for the breakdown of mean effect sizes by each level of the four quality moderators.

Meta-Analysis of Effects From Measures of Language and Comprehension Skills

The meta-regression of standardized language and comprehension measures included 96 effect sizes from 45 studies. As with the foundational skills analysis, the model included the four quality categorical moderators and year of study publication. Results indicated that studies with unacceptable ratings for quality of statistical treatment of data were associated with smaller effect sizes ($b = -0.31$, $SE = 0.18$, $p = .10$). Year of publication also was a statistically significant predictor of smaller effect sizes ($b = -0.02$, $SE = 0.01$, $p = .065$). The I^2 estimate of the percentage of between-study heterogeneity not due to chance variation in effects was 68.03%, with a τ^2 estimate of the true variance in the population of effects of .10. See Table 5 for a breakdown of mean effect sizes and associated parameters by each level of the four quality moderators.

Publication Bias Results

Standardized measures of foundational reading skills. Results of the trim-and-fill analysis indicated that publication bias did not affect

the mean effect size estimate for studies with unacceptable design quality. However, results suggested that unpublished studies with unacceptable quality in the domains of implementation fidelity, statistical treatment, or Type I error and with effect sizes smaller than the mean effect size obtained from published studies are missing from the meta-analysis. See Table 6 for the number of studies likely missing with low quality on each variable and the estimated change to the mean effect size that would result from adding these studies to the meta-analysis. See Figures S3 through S6 in the online supplement for a funnel plot of effect sizes prior to implementing the trim-and-fill procedure for each of the quality areas.

Standardized measures of language and comprehension skills. Results of the trim-and-fill analysis suggested that no studies with unacceptable design quality or unacceptable statistical treatment and effect sizes smaller than the mean effect size from published studies were missing from the meta-analysis. Results also indicated that unpublished studies with unacceptable implementation fidelity quality or unacceptable Type I error likely are missing from the meta-analysis. See Table 6 for the number of missing studies and the estimated changes to the mean effect sizes that would result from including these studies. See Figures S2 through S9 in the online supplement for a funnel plot of effect sizes prior to implementing the trim-and-fill procedure for each of the quality areas.

Discussion

This meta-analysis extended two previous meta-analyses (Wanzek et al., 2016, 2018) both to describe the quality of early reading intervention studies completed between 1995 and 2016 and to investigate how study quality and year of publication were related to the magnitude of effect sizes of supplemental interventions for students with or at risk for reading disabilities in Grades K–3 on standardized measures of foundational reading skills and language and comprehension.

Table 4. Effect Size by Moderator, Standardized Measures of Foundational Reading Skills.

Variable	<i>g</i>	<i>SE</i>	95% CI	<i>p</i>	<i>df</i>	<i>n</i>	<i>k</i>
Design							
Exemplary	.35	.04	[.26, .44]	<.001	38	303	43
Acceptable	.61	.07	[.46, .76]	<.001	28	128	30
Unacceptable	.42	.18	[.03, .80]	.035	14	85	15
Implementation fidelity							
Exemplary	.56	.07	[.40, .71]	<.001	8	69	10
Acceptable	.35	.05	[.24, .45]	<.001	27	222	32
Unacceptable	.50	.07	[.35, .65]	<.001	42	225	44
Statistical treatment							
Exemplary	.42	.06	[.30, .53]	<.001	33	224	36
Acceptable	.44	.10	[.22, .65]	<.001	18	135	19
Unacceptable	.47	.09	[.29, .66]	<.001	31	157	33
Type I error							
Exemplary	.39	.12	[.13, .64]	.005	16	89	17
Acceptable	.48	.07	[.32, .63]	<.001	28	150	31
Unacceptable	.46	.06	[.34, .57]	<.001	37	277	40

Note. CI = confidence interval; *n* = number of effect sizes; *k* = number of studies.

Table 5. Effect Size by Moderator, Standardized Measures of Language and Comprehension Skills.

Variable	<i>g</i>	<i>SE</i>	95% CI	<i>p</i>	<i>df</i>	<i>n</i>	<i>k</i>
Design							
Exemplary	.28	.05	[.17, .40]	<.001	19	51	22
Acceptable	.42	.09	[.24, .60]	<.001	16	29	18
Unacceptable	.20	.27	[-.40, .80]	.47	9	16	10
Implementation fidelity							
Exemplary	.46	.10	[.19, .74]	.007	5	12	6
Acceptable	.27	.07	[.13, .42]	.001	17	40	20
Unacceptable	.36	.10	[.16, .56]	.001	21	44	23
Statistical treatment							
Exemplary	.32	.07	[.16, .56]	<.001	21	41	23
Acceptable	.41	.09	[.22, .60]	<.001	10	33	12
Unacceptable	.22	.16	[-.11, .55]	.18	15	22	16
Type I error							
Exemplary	.33	.13	[.04, .61]	.03	8	15	10
Acceptable	.35	.12	[.09, .61]	.01	14	30	15
Unacceptable	.32	.07	[.18, .46]	<.001	20	51	24

Note. CI = confidence interval; *n* = number of effect sizes; *k* = number of studies.

The Quality of Reading Intervention Studies Published Between 1995 and 2016

Overall, we found that a large percentage of studies received unacceptable ratings on many of the indicators of study quality. Mean quality ratings fell at the unacceptable level for

implementation fidelity, likelihood of Type I error, and statistical analysis. *Implementation fidelity* refers to the degree to which an intervention is delivered as intended. The majority of studies rated as unacceptable on fidelity received the low rating because they did not provide sufficient information to determine the level of implementation of the interven-

Table 6. Publication Bias Analysis Results.

Unpublished Studies with Unacceptable Quality Ratings	Missing studies	Adjusted mean ES	Adjusted 95% CI	Mean ES without missing studies	95% CI without missing studies
Foundational skills measures					
Design	0	NA	NA	.42	[.03, .80]
Fidelity	13	.25	[.20, .31]	.50	[.35, .65]
Statistical treatment	8	.31	[.22, .39]	.47	[.29, .66]
Type I error	15	.22	[.16, .28]	.46	[.34, .57]
Language and comprehension measures					
Design	0	NA	NA	.20	[-.40, .80]
Fidelity	6	.19	[.12, .27]	.36	[.16, .56]
Statistical treatment	0	NA	NA	.22	[-.11, .55]
Type I error	8	.15	[.07, .22]	.32	[.18, .46]

Note. CI = confidence interval; ES = effect size.

tion. Reporting of implementation fidelity data during supplemental reading interventions allows for greater confidence that a significant treatment was in fact due to the intervention being implemented as intended. In addition, implementation fidelity data help researchers know that a null treatment effect was not the result of the intervention not being administered as intended. With implementation fidelity data, educators can look to studies with significant positive effects that were implemented with high degrees of fidelity in order to identify EBPs.

Mean quality ratings fell at the unacceptable level for implementation fidelity, likelihood of Type I error, and statistical analysis.

Sample size and data analysis also affected the quality ratings across studies. Many studies included very small sample sizes (e.g., fewer than 20 students per study group). In addition, some studies used statistical analysis procedures that did not match the design of the study. For example, some studies randomly assigned classrooms to study condition but conducted statistical analysis at the student level. Type I error and statistical analysis are important to investigate as quality indicators, as low-quality ratings in either of these

areas decreases the certainty we have in the findings from a particular study.

The Relationship Between Year of Publication and Effect Size

We found that year of publication predicted effect sizes for both standardized foundational reading skill measures and language and comprehension measures, with more recent studies demonstrating smaller effects. This finding is consistent with those reported in previous research for reading interventions (Scammacca et al., 2015, 2016) but may seem counterintuitive, as one might expect that interventions would demonstrate larger effects over time with the benefit of previous research to identify EBPs with the greatest leverage for improving student outcomes. One possible explanation for effect sizes decreasing as publication year increases is the improvement in the quality of instruction provided in the comparison condition over time, which has been demonstrated in previous research (Lemons, Fuchs, Gilbert, & Fuchs, 2014). Lemons et al. (2014) presented data from five randomized control trials evaluating the efficacy of Kindergarten Peer-Assisted Learning Strategy conducted across a 9-year period. Findings demonstrated a dramatic increase in the performance of comparison students over time, suggesting the need for a more nuanced understanding of how instruction in the counterfactual

affects the identification of EBPs. Most studies did not describe the instruction in the comparison condition in sufficient detail to allow for definitive conclusions to be drawn related to how instruction changed over time.

Another potential explanation for the declining effect sizes over time is that year of publication could be a proxy for global study quality, with more recent studies changing in ways that could account for increases in study quality. This is likely true, as research requirements have become more rigorous than in the past, causing study quality to increase over time (Scammacca et al., 2016). If this is true, year of publication might be demonstrating a small inverse effect for the relationship between overall quality and effect size that could not be attributed to any one specific quality area rating.

The Relationship Between Study Quality and Effect Size

The relationship between study quality and effect sizes was analyzed for both standardized measures of foundational reading skills and standardized measures of language and comprehension. The results indicated that the effect sizes for all four quality variables were significantly different from zero, suggesting that regardless of study quality, supplemental reading interventions had a statistically significant effect in improving reading outcomes. Results also indicated that for standardized measures of foundational reading skills, studies with exemplary design quality were associated with smaller effect sizes. However, with the exception of study design for foundational reading skill measures, variations in study quality indicators were not related to the effectiveness of the supplemental reading interventions in a systematic way. A potential explanation for our inability to detect a systematic relationship between study quality and effect sizes was the variation of quality ratings within studies. A substantial number of contrasts differed in quality ratings even within studies (e.g., sample size differed between treatment groups or fidelity was reported for one treatment but not the other). In addition, a substantial number of studies had different ratings for quality across

the four quality indicators. This variability may have obscured systemic differences in effect sizes based on study quality indicators.

However, with the exception of study design for foundational reading skill measures, variations in study quality indicators were not related to the effectiveness of the supplemental reading interventions in a systematic way.

Publication Bias

Publication bias, or the higher likelihood that studies with positive results would be published compared to studies with low quality and non-significant or very small effects, might have also skewed results related to how study quality relates to effect size. Publication bias likely limited the number of low-quality studies with non-significant or very small effects included in this meta-analysis, limiting our ability to analyze the relationship between effect size and study quality by creating a floor effect. We also explored how publication bias affected the relationship between study quality and the magnitude of effect sizes. We found that a number of unpublished studies with unacceptable quality ratings and small effect sizes were missing from this meta-analysis. We identified evidence of publication bias for three of the four quality indicators for measures of foundational reading skills. On measures of language and comprehension, we identified evidence of publication bias on two of the four quality indicators. Although we cannot conclude that we would have found a more consistent pattern of differences in the magnitude of effect sizes based on study quality if unpublished studies were included, it is important to consider the possibility that publication bias might be responsible for our inability to find a consistent relationship between study quality and magnitude of effect sizes.

Limitations and Future Research

Several factors limit the interpretation of findings. First, as discussed above, the variability

in quality ratings within studies was the most significant limitation affecting our ability to identify a systematic relationship between study quality and effect sizes. Second, the inclusion criteria for this meta-analysis may have limited the pool of studies to those with a minimum standard of quality. The corpus of studies included in this synthesis was limited to studies with experimental and quasiexperimental designs that utilized standardized reading measures, which may have resulted in a corpus of studies that were consistently ranked higher in study quality than would be found in a more heterogeneous group of studies. Finally, due to the statistical analysis being underpowered and the increased risk for Type I error, this meta-analysis utilized a higher alpha level for significance. For this reason, the analysis should be repeated when more studies are available with the $p < .05$ alpha level.

Future research is needed investigating the relationship between study quality and effect sizes of supplemental reading interventions for students with and at risk for reading disabilities. Expanding upon the current meta-analysis to include a wider variety of studies could potentially yield different findings. For example, including studies that utilized less rigorous study designs and unstandardized measures might have resulted in the identification of studies that were more consistently low quality across all four quality indicators, which might have produced different results. In addition, future research might consider including unpublished studies when investigating the relationship between study quality and effect sizes of supplemental reading interventions.

Implications

The purpose of this meta-analysis was to summarize the quality of early reading intervention studies and to analyze the relationship between the study quality and the magnitude of treatment effects of supplemental reading interventions on reading outcomes for students with or at risk for reading disabilities in Grades K–3. Our findings summarizing ratings across each of the quality indicators support the need for researchers to consistently measure and document fidelity of implemen-

tation in research as well as remain cognizant of how to help practitioners implement interventions with high levels of fidelity (Harn, Parisi, & Stoolmiller, 2013). Federal regulations of the Every Student Succeeds Act (2015–2016) mandate the use of evidence-based interventions, or practices and programs that have evidence to show that they are effective at producing results and improving outcomes when implemented. Under ESSA, there are four tiers, or levels of evidence. In order for a supplemental reading intervention to qualify as having strong or moderate evidence, the intervention must be supported by one or more experimental or quasiexperimental studies that had a sufficiently large sample with a similar student population that demonstrated significant, positive effects. In order to help educators identify evidence-based supplemental reading interventions, researchers have a responsibility to implement and report studies with the highest standards of quality to ensure we are recommending interventions based on reliable and valid findings.

In addition to study quality influencing the identification of EBPs, another reason to consider study quality is that it is indicative of the strength and confidence one has in the research findings. Although we found that supplemental reading interventions had positive effects regardless of the level of quality on each indicator, findings from studies with low quality ratings might still be less reliable and valid than the findings from studies with higher quality ratings. For example, studies with small sample sizes are likely less representative of the general population of students with and at risk for reading disabilities and are likely underpowered to reliably detect meaningful effects. In addition, experimental studies that randomize students to condition eliminate potential biases or confounding variables that could impact the effects of a supplemental reading intervention.

Despite guidelines for identifying EBPs that are influenced by study quality, and the knowledge that study quality leads to greater confidence or trustworthiness of findings, our analysis did not identify a systematic relationship between study quality and effect sizes of

supplemental reading interventions. With the exceptions of studies with exemplary study design yielding significantly smaller effect sizes on standardized measures of foundational reading skills, other quality indicators were not related to the effect sizes of supplemental reading interventions. This finding indicates the possibility that study quality might be less important than we hypothesized. We are not yet ready to say with certainty that study quality does not matter until additional research investigates the relationship between study quality and effect sizes utilizing a more heterogeneous corpus of studies; however, if study quality is not related to the effect sizes of supplemental reading interventions, it is worth considering the implication this finding has on future research. For example, there may be a threshold of study quality above which effect sizes are not correlated with further quality, and only meeting these minimum quality standards is needed to identify effective practices. If so, researchers might reconsider spending large amounts of money to conduct research that is far above the threshold of needed quality. In addition, many current syntheses and meta-analyses exclude studies not considered high quality according to current standards. Given our finding that study quality is not systematically related to the effect sizes of supplemental interventions, researchers might reconsider including both low and high quality to represent the full range of research related to a topic.

We are not yet ready to say with certainty that study quality does not matter until additional research investigates the relationship between study quality and effect sizes utilizing a more heterogeneous corpus of studies; however, if study quality is not related to the effect sizes of supplemental reading interventions, it is worth considering the implication this finding has on future research.

In summary, the findings from this meta-analysis demonstrate that overall, a large number of studies received unacceptable ratings on indicators of study quality. In addition, with the exception of study design predicting the effects of foundational reading skills, a systematic relationship between study quality and the effects of supplemental reading interventions was not identified. Supplemental reading interventions were effective regardless of the level of quality on individual indicators.

References

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2013). *Comprehensive meta analysis* (Version 3.3070). [Computer software]. Englewood, NJ: Biostat.
- Carney, K. J., & Stiefel, G. S. (2008). Long-term results of a problem-solving approach to response to intervention: Discussion and implications. *Learning Disabilities: A Contemporary Journal*, 6(2), 61–75.
- Council for Exceptional Children. (2015). *What every special educator must know: Professional ethics and standards*. Arlington, VA: Author.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463.
- Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015–2016).
- Fisher, Z., & Tipton, E. (2015). *robumeta: An R-package for robust variance estimation in meta-analysis*. Retrieved from arXiv:1503.02220
- Fuchs, L. S., Fuchs, D., Compton, D. L., Wehby, J., Schumacher, R. F., Gersten, R., & Jordan, N. C. (2015). Inclusion versus specialized intervention for very-low-performing students: What does access mean in an era of academic challenge? *Exceptional Children*, 81, 134–157. doi:10.1177/0014402914551743
- Goldstein, H., Lackey, K. C., & Schneider, N. J. B. (2014). A new framework for systematic reviews: Application to social skills interventions for pre-schoolers with autism. *Exceptional Children*, 80, 262–286. doi:10.1177/0014402914522423
- Gwet, K. (2001). *Handbook of inter-rater reliability*. Gaithersburg, MD: STATAXIS.
- Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What

- do we really know about fidelity of implementation in schools? *Exceptional Children*, 79, 181–193. doi:10.1177/001440291307900204
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65.
- Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, 43, 242–252. doi:10.3102/0013189X14539189
- Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, 40, 148–182. doi:10.1598/RRQ.40.2.2
- National Center for Education Statistics. (2011). *NAEP: The nation's report card: An overview of NAEP*. Washington, DC: Author.
- National Center for Education Statistics. (2013). *NAEP: The nation's report card: An overview of NAEP*. Washington, DC: Author.
- National Center for Education Statistics. (2015). *NAEP: The nation's report card: An overview of NAEP*. Washington, DC: Author.
- National Early Literacy Panel. (2008). *Executive summary: Developing early literacy. Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy. Retrieved from http://lincs.ed.gov/publications/pdf/NELP_Summary.pdf.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups*. Bethesda, MD: National Institute of Child Health and Human Development.
- O'Connor, R. E., Fulmer, D., Harty, K. R., & Bell, K. M. (2005). Layers of reading intervention in kindergarten through third grade: Changes in teaching and student outcomes. *Journal of Learning Disabilities*, 38, 440–455. doi:10.1177/00222194050380050701
- Pustejovsky, J. E. (2015). *clubSandwich: Cluster-robust (sandwich) variance estimators with small sample corrections*. R package Version 0.0.0.9000. Retrieved from <https://github.com/jepusto/clubSandwich>
- Scammacca, N. K., Roberts, G. J., Cho, E., Williams, K. J., Roberts, G., Vaughn, S. R., & Carroll, M. (2016). A century of progress: Reading interventions for students in Grades 4–12, 1914–2014. *Review of Educational Research*, 86, 756–800. doi:10.3102/0034654316652942
- Scammacca, N., Roberts, G., Vaughn, S., & Stuebing, K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, 48, 369–390. doi:10.1177/0022219413504995
- Swanson, H. L., & Hoskyn, M. (2000). Experimental intervention research for students with learning disabilities: A comprehensive meta-analysis of group design studies. *Advances in Learning and Behavioral Disabilities*, 14, 1–154. doi:10.2307/1170599
- Swanson, H. L., Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes*. New York, NY: Guilford.
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20, 375–393. doi:10.1037/met000001
- Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40, 604–634. doi:10.3102/1076998615606099
- Torgesen, J. K., Houston, D. D., Rissman, L. M., Decker, S. M., Roberts, G., Vaughn, S., . . . Lesaux, N. (2017). *Academic literacy instruction for adolescents: A guidance document from the Center on Instruction*. Portsmouth, NH: Center on Instruction.
- Tran, L., Sanchez, T., Arellano, B., & Swanson, H. L. (2011). A meta-analysis of the RTI literature for children at risk for reading disabilities. *Journal of Learning Disabilities*, 44, 283–295. doi:10.1177/0022219410378447
- Vaughn, S., Wanzek, J., Murray, C. S., Scammacca, N., Linan-Thompson, S., & Woodruff, A. L. (2009). Response to early reading intervention examining higher and lower responders. *Exceptional Children*, 75, 165–183. doi:10.1177/001440290907500203
- Wanzek, J., Stevens, E. A., Williams, K. J., Scammacca, N., Vaughn, S., & Sargent, K. (2018). Current evidence on the effects of intensive early reading

- interventions. *Journal of Learning Disabilities*. doi:10.1177/0022219418775110
- Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*, 541–561.
- Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of Tier 2 type reading interventions in Grades K–3. *Educational Psychology Review, 28*, 551–576.
- Willingham, D. T. (2007). Ask the cognitive scientist: The usefulness of brief instruction in reading comprehension strategies. *American Educator, 30*, 39–45.

Authors' Note

The research described in this article was supported by Grant H325H140001 from the Office of Special Education Programs, U.S. Department of Education; the Eunice Kennedy Shriver National Institute of Child Health and Human Development Grant P50 HD052117; and Grant R324A150269 from the Institute of Education Sciences. Nothing in the article necessarily reflects the positions or policies of the federal government, and no official endorsement by it should be inferred.

Manuscript received January 2018; accepted August 2018.