Journal of Educational Psychology
2016, Vol. 108, No. 5, 609–629

0022-0663/16/$12.00 http://dx.doi.org/10.1037/edu0000082

The Effects of Text Structure Instruction on Expository Reading Comprehension: A Meta-Analysis

Michael Hebert, Janet J. Bohaty, and J. Ron Nelson
University of Nebraska–Lincoln

Jessica Brown
University of Minnesota

In this meta-analysis of 45 studies involving students in Grades 2–12, the authors present evidence on the effects of text structure instruction on the expository reading comprehension of students. The meta-analysis was deigned to answer 2 sets of questions. The first set of questions examined the effectiveness of text structure instruction on proximal measures of comprehension, including examination of potential moderators and effectiveness for students with or at-risk for disabilities. The second set of questions examined the effectiveness on transfer measures of the effectiveness of the intervention across temporal contexts (maintenance), near-contexts (untaught text structures), and far-contexts (general reading comprehension). Overall, the results indicated that text structure instruction improves expository reading comprehension, but the effects were tempered when text structure instruction was compared with stronger comparison groups. The findings also identified 2 moderators that led to increased effect sizes (teaching more text structures and including writing in the instruction). Text structure instruction was also found to be effective across all 3 levels of transfer, although the effects for far-transfer are small and lack consistency. Recommendations include conducting more research to understand the nuances of potential interactions between various instructional approaches and student populations.

Keywords: meta-analysis, text structure, reading comprehension, expository text, informational text

Many children come to school with almost no experience reading expository text (Williams & Pao, 2011) and they have very little exposure to expository text reading in the primary grades (Duke, 2000). This is particularly problematic because reading expository text is a principle way students build the background knowledge necessary to understand content information in the fourth grade and beyond (Saenz & Fuchs, 2002). Furthermore, students who struggle with reading comprehension have particular trouble when reading expository text (Englert & Thomas, 1987; Taylor & Beach, 1984).

The skills needed to read and comprehend expository text are very different than those needed to read and comprehend narrative text (Duke, & Roberts, 2010; Meyer, 1975). Although a conventional set of elements (e.g., characters, setting, plot, solution) can be found across narrative/story grammars (Mandler & Johnson, 1977; Meyer & Rice, 1984), these features are not always found in expository text. Comprehending expository text requires students to make inferences, solve problems, reason, and use complex and varied text structures in ways that are not commonly needed in narrative texts (Armbruster & Anderson, 1980; Snow, 2002). These unfamiliar tasks and structures increase cognitive demands, and may decrease comprehension (Lapp, Flood, & Ranck-Buhr, 1995, 1995; Snow, 2002).

Research suggests teaching students comprehension strategies is an effective way to improve reading comprehension (Duke, Pearson, Strachan, & Billman, 2011). Instruction in expository text structures is another promising approach because text structures are used by authors across a broad range of informational text and aid the reader in organizing facts and ideas in ways that assist retention and recall (Duke & Pearson, 2002; Gersten, Fuchs, Williams, & Baker, 2001; Williams, 2005; Williams & Pao, 2011). Meyer (1975, 1985) described five primary text structures that authors use to present information in expository text; the five structures include description, sequence, cause/effect, compare/contrast, and problem/solution. Authors use compare/contrast to highlight similarities and differences, cause/effect to show a causal relationship, problem/solution to illustrate how a problem was (or might be) solved, description to provide information about the attributes of something, and sequence to group ideas on the basis of order or time.

According to Meyer (1987), text structure instruction might be effective for improving reading comprehension for three reasons. One, knowing the structure of a text helps the reader understand the author's purpose in presenting the information. Two, the reader can use text structures to organize their ideas in order of importance, based on the author's purpose. Three, the reader can save processing time by utilizing the same schema as the author, avoiding the use of valuable cognitive resources searching their memory for an appropriate schema or creating their own. In short, understanding how the author is presenting and organizing information frees up memory and processing resources allowing the reader more capacity to comprehend content (Kieras, 1978).

Michael Hebert, Janet J. Bohaty, and J. Ron Nelson, Department of Special Education and Communication Disorders, University of Nebraska–Lincoln; Jessica Brown, Department of Speech-Language-Hearing Sciences, University of Minnesota.

Correspondence concerning this article should be addressed to Michael Hebert, Department of Special Education and Communication Disorders, University of Nebraska–Lincoln, PO Box 830738, Lincoln, NE 68583-0738. E-mail: michael.hebert@unl.edu

Potential Factors Influencing the Effectiveness of Text Structure Instruction

Researchers have examined two factors that may influence the effectiveness of instruction in text structures: (a) characteristics of students who may benefit from text structure instruction, and (b) instructional approaches and strategies used. The proximity of outcome measures used to assess text structure interventions may also influence the interpretation of their effectiveness.

Student Factors

Students who benefit from text structure instruction may vary from those that do not in a two important ways. First, text structure instruction may be more effective at specific grade levels. Multiple researchers have reported that awareness of text structure increases with age, and instruction in text structures may benefit older students more than younger ones (Englert & Thomas, 1987; Garner et al., 1986; Meyer, Brandt, & Bluth, 1980; Williams, Taylor, & deCani, 1984). However, more recent studies have demonstrated the efficacy of text structure instruction for improving the reading comprehension of students in second grade (e.g., Williams, Hall, & Lauer, 2004; Williams et al., 2005, 2007, 2014; Williams & Pao, 2011; Williams, Stafford, Lauer, Hall, & Pollini, 2009). These more recent findings indicate that text structure instruction may be effective across grade levels.

Second, text structure instruction may be more effective for students of different reading abilities. Studies have shown that students who struggle with reading comprehension do benefit from instruction in reading comprehension strategies and text structures (Gajria, Jitendra, Sood, & Sacks, 2007; Wigent, 2013; Williams, 2005). However, there is evidence that students who struggle with reading comprehension do not appear to benefit as much from this instruction as average or above-average readers (Englert & Thomas, 1987; Swanson, 1999).

Instructional Factors

Instructional factors may also play a role in the effectiveness of text-structure instruction. First, the number and type of text structures taught might influence effectiveness of instruction in text structures. In a recent examination of the history and trends of text structure intervention research Bohaty, Hebert, Nelson, and Brown (in press) noted 68% of the studies conducted on text structures include instruction on one to two text structures, while only 15% of studies examined a more comprehensive approach by teaching all five of the text structures identified by Meyer (1987).

Some text structures may be easier than others to teach and identify. For example, in one study, students understood passages most easily when written with the sequence structure (Englert & Hiebert, 1984). Englert and Hiebert (1984)suggested this might have been due, in part, to students' familiarity with narrative text, in which sequence plays an important role. Other researchers found students' increased awareness and ability to identify certain text structures aided comprehension. For example, Richgels, McGee, Lomax, and Sheard (1987) and Yochum (1991) found that students recalled more ideas after reading the compare/contrast structure than the cause/effect or description structures while Englert and Thomas (1987) found that students more easily identified the sequence structure than the compare/contrast structure.

Second, signal words and graphic organizers are two instructional components often included in text structure instruction. Signal words are those words that indicate or signal the structure of the text. For example, authors may use words like *first, similar, cause,* or *problem* to indicate a sequence, compare/contrast, cause/effect, and problem solution text structure, respectively. The use of signal words may have a positive impact on comprehension partially due to increased coherence, making the passage more logical and well-organized (Siedow & Fox, 1984). However, signal words may be misleading. For example, when comparing and contrasting alligators and crocodiles an author might write, "Crocodiles and alligators differ in three important ways. *First,* crocodiles have more pointed snouts. *Second . . .*" In this case, an inexperienced reader might incorrectly identify this as a *sequence* instead of a *compare/contrast* passage because of the signal words "first" and "second." Additionally, students looking for signal words may ignore content or miss key information necessary for comprehension because they are scanning for signal words.

Graphic organizers are widely used to teach text structures. Although their use as been shown to be effective in a number of studies (Duffy, 1985; Slater, 1988; Ulper & Akkok, 2010; Whittaker, 1992; Wijekumar, Meyer, & Lei, 2012; Wijekumar et al., 2014), they have been less effective in others (Alvermann & Boothby, 1984; Raphael, Englert, & Kirschner, 1986; Walker, 1991; Wilkins, 2007). It may be that graphic organizers are more effective when used in particular content areas (Ciullo, Lo, Wanzek, & Reed, 2016) or prior to rather than after reading (Griffin & Tulbert, 1995). Additionally, some text structures are more naturally organized, decreasing the benefits of graphic organizers. For example, the compare/contrast text structure is organized into two logical, competing arguments. Students who remember one side of the argument are likely to remember the opposing side. Therefore, the graphic organizer may not be advantageous to help recall the information (Alvermann, 1981).

Third, writing is another potential factor that could influence the effectiveness of text structure instruction. In a recent meta-analysis on the impact of writing on reading outcomes, writing was found to have a positive impact on reading comprehension (Graham & Hebert, 2010, 2011). In addition, writing instruction in specific strategies such as text structures improves the overall quality of students' writing (Graham, McKeown, Kiuhara, & Harris, 2012).

Proximity of Outcome Measures to Instructional Approach

Teachers and researchers interested in the effectiveness of text structure interventions may be interested in different outcomes. Some may be interested in proximal outcomes, such as comprehension of specific types of informational text, while others may be interested in how well students transfer those skills to more general reading tasks.

Barnett and Ceci (2002) described researchers' difficulties deciding how to label transfer (i.e., near, far), and provided direction for future researchers to label transfer. The authors described knowledge as being measured on a continuum from near transfer to far transfer across six dimensions: (a) the knowledge domain, (b) physical context, (c) temporal context, (d) functional context, (e) social context, and (f) modality (Barnett & Ceci, 2002). When examining the effectiveness of instructional approaches, research-

ers are often implicitly concerned with characterizing transfer of knowledge across time and outcomes (e.g., Graham & Hebert, 2010; Kim, Vaughn, Wanzek, & Wei, 2004; Piasta & Wagner, 2010). To understand the effectiveness of text structure instruction, it is important to specify and categorize measures, and examine the proximity of the measures to the instructional approach.

The Current Study

A meta-analysis is needed to determine the effectiveness of text structure instruction across a range of participants and contexts. Although several reviews of this literature have been conducted (Bohaty, Hebert, Nelson, & Brown, in press; Meyer, 1979; Meyer, 1987; Meyer & Ray, 2011; Slater, 1988), no meta-analytic review of this literature has been conducted to date. Such a meta-analysis of the literature may provide a more accurate estimate of the true effect of text structure instruction by creating an average weighted effect size across studies, including studies that have not been published, as well as provide an analysis of potential moderator variables.

To examine the impact of text structure instruction (TSI) on expository reading comprehension across studies, we conducted a meta-analysis to answer the following sets of research questions (RQs) for students in Grades 1 to 12:

1. (RQ1): Does TSI improve students' comprehension of expository text on proximal measures of comprehension?
    a. (RQ1a) To what extent do the results of an aggregated ES approach to calculating the average weighted ES compare with results of robust variance estimation?
    b. (RQ1b): Are the effects of TSI moderated by factors related to instructional approaches, participants, or outcome measures used?
    c. (RQ1c): Is TSI effective for students with or at-risk for disabilities?
2. (RQ2): Do the effects of instruction in expository text structures transfer to distal measures of comprehension?
    a. (RQ2a): Are the effects of TSI maintained over time (temporal transfer)?
    b. (RQ2b) Does TSI improve comprehension of text written in an untaught structure (near-transfer)?
    c. (RQ2c) Does TSI improve general reading comprehension outcomes measured by norm-referenced tests (far-transfer)? (Note: We classify norm-referenced tests, or standardized tests, of reading comprehension as "general reading comprehension" because the tests usually include items that sample broadly across the target domain.)

Method

Our research questions provided the basis on which we developed our initial inclusion and exclusion criteria, as well as the foundation for our coding and analysis. At times, studies were collected that tested the limits of our initial plans for our criteria, coding schemes, and analyses; we refined them accordingly. However, all decisions were anchored by our guiding research questions.

Inclusion and Exclusion Criteria

The strategies used for locating and selecting studies for inclusion were influenced by 11 factors. If a study's author did not include enough information to make a determination of whether it met the criteria, an attempt was made to contact the author. If the author could not be contacted or provide the necessary information, the study was eliminated, ensuring that the included studies reflected the intent of the review.

We deemed studies to be eligible for inclusion if they met the following criteria:

1. The researchers provided empirical evidence relevant to the research questions.
2. The report was published in English (we did not translate reports).
3. An experimental, quasiexperimental, or counterbalanced design was employed. The investigators of the study had to establish equivalence of the experimental and comparison groups for quasiexperimental studies.
4. The study was conducted with school-age participants in Grades 1 through 12.
5. Students in the treatment group received instruction in one or more of five expository text structures identified by Meyer (1985): *description, sequence, compare-contrast, cause-effect,* and *problem-solution*. Instruction was defined as reading or writing instruction in which the students were taught something about how to use text structures to improve their comprehension (e.g., how to identify the text structure, how to answer questions about text structure, how to construct a written text in one or more of the text structures).
6. An expository reading comprehension outcome measure, or norm-referenced measure of reading comprehension, was included. Researcher-created measures were acceptable.

We excluded studies if

7. Students in the control group received TSI.
8. Hierarchal structure of text was examined instead of the expository text structures identified by Meyer (e.g., Taylor, 1982).
9. Students in the treatment group received only an overview of text structure, without instruction. For example, we eliminated studies in which students received graphic organizers depicting text structures, but were not taught how to use the graphic organizers (e.g., Brandt, 1978). In

these cases, we decided that students were provided with tools, but not instruction. Similarly, we did not include studies comparing students who received text with inserted questions about text structure with students who received text without the questions (e.g., Ordynans, 2012); this was also not considered instruction, as it did not teach students a strategy they could independently implement.

10. The treatment group received additional reading instruction that differed from the control group, above and beyond TSI (e.g., Ward-Washington, 2001). Due to confounding factors in these cases, it would have been impossible to determine whether any effects were due to TSI, the additional reading instruction, or a combination of the two.

11. The researchers did not include data necessary to calculate an average weighted effect size.

In addition,

12. To be included in the maintenance analysis, studies were required to have a measure of reading comprehension that occurred at least one day after the posttest.

13. To be included in RQ2a or RQ2b, studies were required to include a measure of reading comprehension for text written using an untaught structure or a norm-referenced measure of general reading comprehension, respectively.

Search Strategies Used to Locate Studies

A search that was as broad as possible was undertaken to identify relevant studies for this review based on the inclusion criteria. Ninety-nine electronic searches were run across six databases (i.e., ERIC, Education Index Retrospective, PsycINFO, Academic Search Premier, ProQuest—including Dissertation Abstracts International, and Web of Science) to identify relevant studies with electronic records through January 2014. For the 3,121 items identified through the electronic searches, two authors read each entry. If the item looked promising based on its abstract or title, it was obtained. Once a document was obtained, the reference list was searched to identify additional possible studies for inclusion. Among the 307 documents collected, 45 manuscripts included studies that met the inclusion criteria.

Two authors read every study and collectively decided which studies met criteria for inclusion and exclusion. Reliability of inclusion and exclusion of studies was then conducted by an independent rater. For reliability purposes, a third author examined a randomly ordered set of all the included studies and a sample of the excluded studies using the inclusion and exclusion criteria. Percentage of total agreement was 95%. There were two disagreements, which were resolved through discussion.

Categorizing Studies According to Questions and Methods

Each study was read and placed into a category based on the question it answered. For RQ1, we included studies comparing TSI with any comparison condition, including business-as-usual (BAU), no treatment, or competing comprehension instruction. If a study included multiple comparison conditions (e.g., a BAU condition *and* a competing comprehension instruction condition), we chose to include both conditions in the analysis using robust variance estimation (RVE) to account for statistical dependencies between conditions related to correlated effects (see Tanner-Smith & Tipton, 2014).

Studies included for RQ1 were used in the moderator analysis for RQ2 and further examined and placed into subcategories based on whether they included data relevant to RQ3 (i.e., maintenance of the effects over time, transfer to a new structure, or general reading comprehension measured by norm-referenced tests). All studies were subsequently reexamined to verify that they were included for the appropriate question.

Study Feature Coding

Each study was coded for variables in three categories: study descriptors, quality indicators, and variables necessary to calculate effect sizes. Study descriptors and quality indicators were chosen to contextualize this report for external validity and/or for their potential to account for variability in the average weighted effect sizes. Study descriptor variables included: grade level, type of student (e.g., struggling readers, English Language learners, etc.), number of participants, study locale, treatment setting, treatment length, training of participants, description of the treatment, description of the control condition, subject area, text structure(s) studied, genre, outcome measures, publication type, research design, and random assignment.

Study quality indicators were chosen to evaluate studies based on standards for design, implementation, and measurement. The quality indicators included: randomization with analysis at the appropriate level, total attrition of less than 10%, equal attrition across groups (within 5%), well-defined control groups, controls for teacher effects, more than one teacher per treatment group, reported fidelity, reliability of the measures of greater than 60%, and no ceiling or floor effects on posttests (contact the first author for an expanded definition of the quality variables). Because quasiexperiments were required to have a pretest to be included, additional indicators were necessary for those studies (i.e., no ceiling or floor effects on pretests, and equivalence of groups on pretest measures).

An overall quality score was calculated for each study based on the coded quality indicators. Each study was awarded one point for meeting the criterion for each quality indicator, with the exception of quality indicators related to the outcome measures. For studies that employed multiple outcome measures, quality indicators related to the measures were scored as the percentage for which the study met the quality criteria. To make the quality scores comparable across both types of studies, the proportion of total points possible for each study was used.

Two authors coded all of the studies. Percentage of total agreement was 91.9%. Disagreements were resolved through discussion.

Calculation of Effect Sizes

Effect sizes calculated for this review were based on expository reading comprehension outcome variables and norm-referenced

reading comprehension outcomes. Effect sizes were calculated for all reading comprehension measures in each study, and all measures were reported as continuous variables. Hedge's *g* was used to represent intervention effects on outcome measures identified for each study to provide an unbiased effect for each study, including those with small samples (Hedges, 1981, as cited in Lipsey & Wilson, 2001).

Hedges (2009) recommends meta-analysts choose a model to estimate or adjust effect size calculation parameters in such a way that they are consistent and analogous to effect sizes of other studies to which the study will be compared. Based on the studies located for this review, effect sizes were calculated and adjusted in three ways (specific equations and calculation information are provided in Appendix A). One, for true-experiments with randomization and data analysis at the student-level, the standardized mean difference effect size (*d*) was calculated and the small sample correction was applied (Hedges, 1981, as cited in Lipsey & Wilson, 2001). Two, for true experiments with randomization and data analysis at the cluster-level, a cluster-level effect size was calculated based on the between-groups variance and then multiplied by the intraclass correlation using methods described by Hedges (2009), prior to applying the small sample correction. Because the studies included in this analysis ($k = 6$) did not report the intraclass correlations, they were imputed at .20, based on the conventions of the What Works Clearinghouse Procedures and Standards Handbook (What Works Clearinghouse, 2014).

Three, quasiexperiments (e.g., assignment at the group- or classroom-level, but analysis at the participant-level) were required to include a pretest, due to potential assignment bias. We adjusted the ESs for pretest differences between groups when calculating the ESs (*d*). As with the cluster-level experiments, the quasiexperiments in this review ($k = 24$) did not report appropriate data to calculate classroom-level variance, requiring us to impute the intraclass correlation. We imputed the intraclass correlation at .20 and adjusted the effect sizes using by intraclass correlation estimator "ES = d_T" (Hedges, 2009), before applying the small sample correction.

Additional effect size considerations. Some studies identified for inclusion in this review did not report statistics in the form of means and standard deviations. When possible, available data was converted into a form usable for calculating an effect size. For example, missing standard deviations were estimated from summary statistics reported by researchers or by estimating residual sums of squares to compute a root mean squared error (RMSE; e.g., Shadish, Robinson, & Congxiao, 1999). For one study (Alvermann, 1982), means were estimated from graphic data, while the standard deviations were estimated from another study (Alvermann, 1981) in which the researcher used an identical measure with a similar sample (Lipsey & Wilson, 2001).

Winsorizing overly influential studies. After calculating an effect size for each of the studies, we examined the distribution of effect sizes to look for potential outliers in the meta-analyses conducted for each of the research questions. To ensure extreme effect sizes did not have a disproportionate influence on the average weighted effect size, we Winsorized (see Lipsey & Wilson, 2001) the effects by recoding the outliers at three times the interquartile range of the sample. The set of study effect sizes for researcher created measures pertaining to RQ1 included no effect sizes that needed to be Windsorized. The TSI versus general reading approaches subset of study effect sizes for RQ1c (maintenance) included one outlier (i.e., Bakken et al., 1997, ES = 3.11), which was recoded to an effect size of 1.66. There were no outliers found in the distribution of effect sizes for RQ2a or RQ2b.

Statistical Analysis of Effect Sizes

Random effects models were used for all analyses, as the intent of this review was to generalize beyond the population of studies in this analysis. For each analysis, we calculated the mean and confidence intervals for weighted effect sizes. The smallest number of studies included in an analysis was five. Similar to conventions of other meta-analyses, we decided to conduct meta-analyses only for treatments that contained four or more independent comparisons assessing the same reading construct (Graham & Hebert, 2010; Graham & Perin, 2007).

Like much educational research, many of the studies in the review included data that could be used to calculate multiple effect sizes within a single study. Three such situations occurred in this review: (a) *multiple treatments* with a single control group, (b) *multiple subgroups* within a treatment, and (c) *multiple outcome measures*. Scammacca, Roberts, and Stuebing (2014) outlined multiple ways to account for dependencies in meta-analyses related to multiple effect sizes, including selecting a single outcome, aggregating all measures within a study, a shifting-unit-of-analysis approach, calculating effect sizes for studies with multiple measures by incorporating the correlation between measures, combining treatment groups, conducting a three-level meta-analysis, or use of RVE. In the current study, we elected to use two approaches. Our primary approach was RVE because it allowed us to include multiple effect sizes per study. However, there were instances in which we supplemented the approach by aggregating measures with a shifting-units-of-analysis approach due to the limitations of RVE and for sensitivity testing.

Use of robust variance estimation (RVE) for RQ1 and RQ2. The RVE approach can be applied no matter the source of dependence among the effect sizes (Hedges, Tipton, & Johnson, 2010; Scammacca et al., 2014; Tanner-Smith & Tipton, 2014). Use of the approach allows the meta-analyst to include multiple individual effect sizes from a single study, without the need to combine groups, average effect sizes within studies, or combine groups.

The specific RVE approach used for this study was the "correlated effects case" described by Tanner-Smith and Tipton (2014) in their tutorial. Average weighted effect size calculations and metaregression models were conducted in Stata/SE 12.1 using the robumeta.ado macro suggested by Tipton and Tanner-Smith (2014) and obtained for Stata/SE version 12.1 directly from Hedberg (E.C. Hedberg, personal communication, March 26, 2015). A between-study correlation of 0.80 was imputed. Sensitivity analyses were then conducted to examine the impact of the correlation (see Tipton & Tanner-Smith, 2014).

Metaregression analyses. Metaregression models were used to examine the potential moderating effect of other study characteristics. The number of moderator variables used was limited by the degrees of freedom (based on the number of studies, not the number of effect sizes). Although studies with multiple comparisons were used, these studies each only contributed one degree of freedom to the metaregression model. However, RVE allowed for examination of moderator variables that varied both between- and

within-studies, providing a more thorough analysis. All within-study moderator variables (i.e., *competing treatment, grade level, number of sessions*) were centered around the variables' study means to estimate the within-study effects.

The variables of interest for moderator analyses were chosen *a-priori*, as discussed in the introduction. In most cases, moderators were examined using metaregression. For one variable, *student ability*, it was impossible to completely parse out the effects for students with disabilities.

Analyses to supplement the use of RVE. Despite the advantages of using RVE for dealing with dependencies in the data, the approach is still emerging. Only three reports were located that used this approach in educational research (i.e., Scammacca et al., 2014; Uttal et al., 2013; Wilson, Tanner-Smith, Lipsey, Steinka-Fry, & Morrison, 2011). Thus, there is a lack of clarity on some issues pertaining to new analysts who wish to use this approach. For example, the Stata 12.1 macro developed for this approach does not calculate a Q statistic or an I^2 statistic for the group of studies, which are traditionally used for reporting heterogeneity (the developer of the Stata macro is currently working on adding the I^2 statistic to the output but this was not completed before the current review (E.C. Hedberg, personal communication, March 26, 2015). Instead the RVE macro for Stata 12.1 calculates a τ^2 statistic that provides a method of moments between studies variance component that accounts for dependent effect size estimates. Because the Q and I^2 statistics have been more traditionally used to determine whether enough heterogeneity is present to warrant a metaregression analysis, we used a shifting-unit-of-analysis approach with aggregated effect sizes within studies in coordination with the RVE to compare the methods. Publication bias analyses are also not straightforward when using the RVE approach; it does not seem appropriate to include multiple effect sizes per study due to the fact that the publication bias macros in Stata 12.1 treat each effect size as an independent study.

Aggregated effect sizes and shifting-units-of-analysis approach. We aggregated effects within studies to report an overall effect size for each study, conduct the publication bias analysis, and provide the Q and I^2 statistics. This also allowed us to calculate an average weighted effect size that could be compared to the effect size found using the RVE approach for the purposes of sensitivity analysis. To aggregate the effect across groups on one measure within a study, we used techniques outlined by Cortina and Nouri (2000) (see Appendix A for the equations and explanation). Across measures, we simply averaged the effect sizes.

In addition to sensitivity analyses, we had to use an aggregated approach in the analyses related to RQ2 because there were either (a) too few effect sizes to use RVE (i.e., RQ2b and RQ2c); or (b) too few effect sizes per study for a moderator analysis (i.e., RQ2a; see Tanner-Smith & Tipton, 2014). Therefore, all analyses related to research question 2 were analyzed with aggregated effect sizes using a shifting-unit analysis approach.

Breakout analysis for studies which included only students with or at-risk for learning disabilities. Another question asked in this study was whether this approach was effective for students with disabilities. However, we could not parse out the effects for these students in some studies, because many did not include information on the outcome measures specific to these students. Therefore, we elected not to include this variable in the metaregression. Instead, we conducted a separate analysis with studies that included only students with or at-risk for learning disabilities. This allowed us to estimate the effects of TSI for students with or at-risk for disabilities while not contaminating the effects with potential comparisons involving some similar students. For studies in which we could parse out the effects for subsets of students, we did so (i.e., Ulper & Akkok, 2010).

Sensitivity and Publication Bias Analyses

Some estimation was required in the calculation of the data used in this meta-analysis. For example, some studies reported the length of the intervention in weeks or months, rather than in sessions or minutes; due to a lack of precision in estimating minutes, we imputed an average of two sessions per week to calculate an approximate total number of sessions in these instances ($k = 8$). In other cases, we had to impute intraclass correlations and/or the within-study between effects correlation for the RVE analysis. Sensitivity analyses were performed on analyses involving imputed missing values. Additionally, the possibility of publication bias was assessed using funnel plots, Egger's test for small study effects, and an exploratory trim and fill analysis.

Results

Overall, the literature search yielded 45 reports, from which 323 effect sizes were extracted. A summary of the descriptive, participant, and treatment characteristics for the 45 studies are presented in Table 1. Forty of the 45 studies provided information related to RQ1; the additional five studies included only norm-referenced measures of general reading comprehension, so they were used in RQ2c only. Within the set of studies for RQ1, we identified 14 studies with maintenance measures (RQ2a) and seven studies with measures examining the transfer of the effects of instruction to an untaught structure (RQ2b). Nine total studies were found examining the transfer of effects of instruction to norm-referenced reading comprehension outcomes (RQ2c); four of the studies included in RQ1 and the five additional studies that did not include proximal measures.

A summary of the information on individual studies, including study characteristics, treatment characteristics, an overall study quality score, and the overall effect size for each study can be found in Table 2. The overall effect size represents an aggregation of the effect sizes across measures and subgroups within treatment conditions, but each treatment group comparison is presented separately due to the inclusion of some of the same participants in each comparison.

The percentage of studies that met the quality criteria for each quality variable across the set of studies can be found in Table 3. Quality was scored for the set of studies used in RQ1a and RQ2c, as the set of studies in RQ2c included four studies used in RQ1 as well as five additional studies that did not meet the inclusion criteria for RQ1a. Across the studies included, reliability of the study measures was reported most consistently (80% and 100% in RQ1a and RQ2c, respectively), while fidelity was reported least consistently (20% and 13% in RQ1a and RQ2c, respectively).

We calculated bivariate correlations to examine the relationships between some of our study variables and the effect size. The correlations can be found in Table 4. Examination of the relationship between study quality (research implementation quality) and effect size was nonsignificant and essentially zero. In fact, none of the

Table 1
Summary of Descriptive, Participant, and Treatment Characteristics for All Reports

Characteristic	*k*	%
*Publication year		
1970s	1	2
1980s	14	31
1990s	8	18
2000s	14	31
2010–present	8	18
*Form of publication		
Journal	21	47
Dissertation	20	44
Conference paper	3	7
Other	1	2
School level (Grades)		
Elementary (1–5)	21	47
Secondary (6–12)	22	49
Length of instruction		
1 session	5	11
2–5 sessions	7	16
6–10 sessions	11	24
11–20 sessions	10	22
More than 20	12	27
Structures taught		
One	13	29
Two	15	33
Three	4	9
Four	5	11
Five	8	18
Study type		
Experiment (SL)	15	33
Experiment (CL)	6	13
Quasiexperiment	24	53
Effective sample size		
25 or less	16	35
26–50	13	29
51–100	7	16
>100	9	20
Student type		
Ss w/LD or at-risk	7	16
Ss not at-risk or FR	38	84
Locale		
Urban	12	27
Suburban	13	28
Rural	4	9
Multiple locations	4	9
Cannot tell	12	27
Signal words		
Yes	20	44
No	25	56
Writing		
Writing	29	64
No writing	16	36
Graphic organizers		
Yes	34	76
No	11	24

Note. CL = cluster level; FR = full-range; *k* = number of studies; LD = learning disability; SL = student level; Ss = students
* Based on the published or most recent report found.

potential moderator variables were significantly correlated with the effect size. Despite this, some variables included in the moderator analysis for RQ1 *were* significant predictors of the effect size (presented in the results section for RQ1b). This is primarily a result of the difference between the effect sizes used in the correlational analysis and the effect sizes used in the RVE meta-analysis. Specifically, the correlational analysis was conducted using the aggregated effect sizes for each study, so as not to overweight the contribution of studies with multiple effect sizes as compared with studies with only a single effect size in the correlations. On the other hand, the meta-analysis and subsequent moderator analysis were conducted using multiple effect sizes for each study, using RVE to account for dependencies among effect sizes within each study.

Summary of the Results for Research Question 1 and Subquestions: Effects of TSI on Proximal Measures of Comprehension

The studies meeting the criteria for RQ1 often included data for the calculation of multiple effect sizes. As mentioned in the Method section, we conducted the meta-analysis of these effect sizes in two ways (RVE and using one aggregated ES per study) to examine the similarities and differences between the approaches (RQ1a). We organized this section into subsections to: (a) compare the results across the two analyses; (b) present the metaregression analysis of potential moderators using RVE (RQ1b); and (c) present a breakout analysis of TSI using aggregated ESs for students with or at-risk for learning disabilities (RQ1c).

RQ1a: Comparing results using a single aggregated effect per study with results using RVE. The results comparing the aggregated effect size approach and the RVE approach can be found in Table 5. Forty studies with data relevant to these analyses yielded 170 effect sizes. All of the 170 effect sizes were used in the RVE analysis. The results were calculated using an assumed within-study between effect sizes correlation of .80. Sensitivity analyses were then conducted using different within-study correlation values between .01 and .99, but the effect sizes and confidence intervals were unchanged across all of the values (see Appendix B). The average weighted effect size across the 45 studies was 0.57, 95% CI [0.39, 0.76] in the RVE model.

Although 170 effect sizes were used in the RVE analysis, only 118 of the effect sizes could be aggregated within studies to arrive at one ES per study ($k = 40$). The additional effect sizes could not be used in the aggregated analysis because they included comparisons of multiple treatments that involved the same participants. For those studies, we were forced to choose only one of the treatment comparisons; we chose effect sizes comparing treatment to business-as-usual conditions or the weakest comparison (the effect sizes used are superscripted in Table 1). The average weighted effect size across the 45 studies was 0.56, 95% CI [0.43, 0.69] in the aggregated model.

The comparison of RVE to the aggregated effects model reveal that the results of the RVE analysis are very similar to the results of the aggregated analysis. However, the RVE analysis has a wider confidence interval. This does not influence the inferences that were made regarding RQ1a because the confidence intervals for both analyses were not close to crossing zero. The use of RVE allowed us to utilize all of the available information from the studies in our subsequent moderator analysis.

The *Q*-statistic from the aggregated model indicated more heterogeneity in the sample than would be expected from sampling error alone, as the *Q*-value of 118.60 was almost double the critical value of 54.57 for a chi-square distribution with 39 degrees of freedom at the .05 significance level. Further, the I^2 for the sample indicated that

Table 2
Descriptive and Effect Size Information for All Studies

Study	Study type	Grade	Students	Quality score	Instruct	Signal words	GO	Writing	Structures taught (# sessions)	Comparison group[a]	# of ESs[b]	Overall posttest ES	Transfer ESs
Alvermann (1981)	E	10	FR	.67	0	No	Yes	—	SD, CC (1)	Content	2	.62[c]	.91[d]
Alvermann (1982)	E	10	A & AA	.44	0	No	Yes	—	SD, CC (1)	Content	2	1.62[c]	
Alvermann & Boothby (1984)	E	4	FR	.67	1	No	Yes	—	SD (14)	Content	1	.74[c]	
									SD (7)	Content	1	−.84	
Bakken et al. (1997)	E	8	LD	.55	1	Yes	No	S	SD, SQ (3)	Content	3	2.17[c]	1.54[d]
										CT	3	.46	2.22[e]
Bartlett et al. (1980)	Q	5	FR	.36	0	No	No	P	SD, CC, PS, CE (15)	BAU	1	1.17[c]	.14[d]
Bartlett (1978)	Q	9	No EBD	.36	1	No	No	P	SD, CC, PS, CE (5)	BAU	1	.08[c]	
Bohaty (2015)	E	4 & 5	LD	.75	1	No	No	—	All 5 (8)	BAU	3	.62[c]	−.17[f]
											1		
Brimmer (2004)	Q	6	FR	.73	0	Yes	No	—	All 5 (12)	CT	1		−.10[f]
Broer et al. (2002)	Q	6	FR	.66	1	Yes	Yes	—	CE (16)	Content	2	.21[c]	
Coleman (1983)	E	9	A	.44	0	No	Yes	—	SD, CC (1)	Content	1	.16[c]	.20[d]
Crowe et al. (2014)	E	1, 2, 4	<97%ile	.78	1	Yes	Yes	—	CC, SQ, CE (48)	BAU	6		−.02[f]
Duffy (1985)	E	6	FR	.44	0	Yes	Yes	N	All 5 (35)	BAU	1	1.37[c]	
Englert et al. (1991)	Q	4 & 5	FR	.36	1	Yes	Yes	N & P	SQ, CC (~40)	BAU	2	.12[c]	
Gentry (2006)	E	4	FR	.56	1	No	Yes	N & S	All 5 (5)	CT	1		.10[f]
Gould (1987)	Q	4–8	FR	.18	0	Yes	Yes	—	SD, CC, PS, CE (8)	BAU	40	.20[c]	
Hall et al. (2005)	E-CL	2	FR	.59	0	Yes	Yes	P	CC (~15)	BAU	2	1.02[c]	−.05[e]
										Content	1	1.49	
Hammann & Stevens (2003)	Q	8	A & AA	.41	1	Yes	No	P	CC (6)	BAU	2	.41[c]	
									CC (6)	CT	2	.56	
									CC [+AT] (6)	CT	2	.29	
Hickerson (1986)	Q	7 & 10	No LD	.45	1	Yes	Yes	N & P	ALL 5 (~12)	BAU	2	.84[c]	
Hoffman (2010)	Q	5	FR	.59	1	No	Yes	—	CC (8)	Content	1		−.18[f]
McDermott (1990)	Q	4	No LD	.45	0	No	Yes	P	SD, PS (10)	BAU	2	.85[c]	
McLaughlin (1990)	Q	5	AR	.78	0	No	Yes	P	CC (1)	CT	2	.11[c]	
Meyer et al. (2002)	E	5	No LD	.63	1	Yes	Yes	P	CC, PS (30)	BAU	2	.35[c]	.74[d]
Moore (1995)	Q	6	A & AA	.64	1	Yes	Yes	P	CC, CE (~14)	Content	1	.72[c]	.40[d]
													.52[e]
Newman (2007)	Q	3	FR	.45	1	No	Yes	P	SD, CC, SQ (24)	Content	2	1.32[c]	
Occasio (2006)	Q	5	AR	.45	1	Yes	No	P	CC, SQ, PS, CE (16)	CT	1	2.81[c]	
Raphael et al. (1986)	Q	5 & 6	No LD	.18	1	Yes	Yes	P	CC, PS (~40)	BAU	2	.66[c]	
									CC, PS (~40)	CT	2	.06	
									CC, PS[+AT] (~40)	CT	2	−.16	
Reese (1988)	E	9	A & AA	.56	1	No	Yes	—	All 5 (10)	BAU	1		.16[f]
Reynolds & Perin (2009)	Q	7	A	.82	1	No	Yes	P	SQ (5)	Content	2	.68[c]	1.11[d]
										CT	4	.11	.89[e]
Russell (2005)	Q	9	AR	.55	1	No	Yes	P	SD, CC, SQ, CE (9)	BAU	1	1.96[c]	
Samson (1982)	Q	9–11	CB	.49	1	Yes	No	N	SD, CC (8)	BAU	1	−.56[c]	
Scott (2011)	Q	6	FR	.45	1	No	Yes	P	SD, CC, SQ (81)	Content	1	.56[c]	
Slater et al. (1985)	E	9	FR	.78	0	No	Yes	—	PS, CE (1)	Content	2	.29[c]	
										CT	2	.46	
Smith & Friend (1986)	Q	9–12	LD	.33	1	Yes	No	—	ALL 5 (4)	CT	1	.96[c]	.75[d]

Table 2 (*continued*)

Study	Study type	Grade	Students	Quality score	Instruct	Signal words	GO	Writing	Structures taught (# sessions)	Comparison group[a]	# of ESs[b]	Overall posttest ES	Transfer ESs
Spires et al. (1992)	E	4	FR	.61	0	Yes	No	—	CC, PS (6)	Content	12	.39[c]	.26[d]
										CT	12	−.44	
Taylor (1985)	E	6	FR	.61	1	No	No	P	CC (10)	Content	1	−.38[c]	−.07[d]
									CC (10)	CT	1	−.52	
									CC [+AT] (10)	CT	1	−.01	
Ulper & Akkok (2010)	Q	6	AR & AA	.55	0	No	Yes	P	PS (2)	BAU	1	.94[c]	
										CT	1	.10	
Walker (1991)	Q	5	FR	.36	1	Yes	Yes	—	All 5 (~7)	Content	1	.09[c]	−.15[d]
									All 5 (~7)	CT	1	−.34	
									All 5 [+AT] (~7)	CT	1	−.09	
Whittaker (1992)	Q	6	A & AA	.45	0	No	Yes	—	SD, CC (12)	BAU	2	.20[c]	
Wijekumar et al. (2012)	E-CL	4	FR	.71	1	Yes	Yes	S	CC, PS (~26)	BAU	2	.20[c]	.13[f]
											1		
Wijekumar et al. (2014)	E	5	FR	.71	1	Yes	Yes	S & P	CC, PS (~26)	BAU	2	.31[c]	.25[f]
											1		
Wilkins (2007)	Q	7 & 8	AR	.64	1	Yes	Yes	S	CE (5)	CT	1	−.07[c]	−.25[e]
										CT	1	−.56	
Williams et al. (2005)	E-CL	2	FR	1.00	1	Yes	Yes	S	CC (15)	BAU	2	.78[c]	−.18[e]
										Content	1	.78	
Williams et al. (2007)	E-CL	2	FR	.89	1	Yes	Yes	S	CE (22)	BAU	3	.41[c]	
										Content	3	.44	
Williams et al. (2009)	E-CL	2	FR	1.00	1	Yes	Yes	P	CC (22)	BAU	6	1.35[c]	.93[e]
										Content	2	.84	
Williams et al. (2014)	E-CL	2	FR	.78	1	Yes	Yes	S	CE (22)	BAU	3	.75[c]	
										Content	3	.60	

Note. A = average; AA = above average; AR = at-risk students; BAU = business as usual; CB = college bound; CC = compare/contrast; CE = cause/effect; CT = competing treatment; E = experiment; E-CL = true experiment at the classroom level; ES = effect size; FR = full-range of classroom abilities; GO = graphic organizer; LD = learning disability; N = notes; P = paragraph; PS = problem/solution; Q = quasiexperiment; S = sentence; SD = simple description; SQ = sequence.

[a] Content = Students learned or read the same content at the treatment group but were not taught a specific intervention, while alternative treatments involved instruction designed to improve comprehension other than text structure instruction. [b] Number of effect sizes used in the RVE analysis for RQ1a. [c] Effect size used in aggregated analysis for RQ1a. [d] Effect size for temporal transfer in RQ2a. [e] Effect size used in analysis of near-transfer measures in RQ2b. [f] Effect size used in analysis of standardized measures of general reading comprehension for RQ2c.

Table 3
Proportion of Studies Meeting Each Quality Standard

Quality feature	RQ1a (k = 40)	RQ2c (k = 9)
Randomization w/analysis at correct level	.45	.75
Total attrition <10%	.64	.75
Equal attrition across conditions	.61	.75
Control condition defined	.39	.38
Fidelity reported	.20	.13
Teacher effects controlled for	.47	.67
More than 1 teacher per group	.41	.33
Reliability of the measure >.60	.80	1.00
No posttest ceiling or floor effects	.74	1.00
*No pretest ceiling or floor effects	.80	1.00
*Pretest equivalence	.80	1.00

Note. k = number of studies; RQ1a = Research Question 1a; RQ2c = Research Question 2c.
* Standards applied only for quasiexperimental studies.

67% of the variance was due to between study factors in the aggregated model. We would expect even more heterogeneity to be present with the inclusion of additional effect sizes for the RVE. Indeed, the comparison of the τ^2 statistic was larger based on the RVE model (0.15 compared with 0.07 in the aggregated model). Based on these factors, we rejected the null hypothesis of homogeneity within the sample, instead attributing the differences in effect sizes in the sample to something other than subject-level sampling error. Thus, we conducted a moderator analysis in an attempt to explain some of the variation between studies.

RQ1b: Moderator analysis for the effects of TSI. As previously stated, the moderator analysis was conducted using metaregression with robust standard error estimation. The estimate of between-effect within-study variation (ρ) used for the analysis was .80. Sensitivity analyses indicated the findings were consistent across different values of (ρ). The number of moderators included in the model was limited by the number of studies included, not by the number of effect sizes (Tanner-Smith & Tipton, 2014). Thus, we limited our moderator analysis to six moderators.

The six moderators included in the analysis were: (a) a dummy variable comparing BAU to a competing treatment (this variable was centered due to the variation with studies); (b) a dummy variable comparing studies involving elementary school participants to middle and high school participants; (c) a continuous variable indicating the number of text structures (1–5) taught in the study (this variable was coded 0–5 to aid in the interpretability of the intercept, as it was not possible for a study in this analysis to include instruction in fewer than one text structure); (d) a dummy variable indicating whether a variation of explicit instruction was used for instruction; (e) a dummy variable indicating whether writing was included in the instruction; and (e) a dummy variable indicating whether students were taught signal words during instruction. We chose these variables because they represented information about whether the effect size varied within or across studies as a function of the strength of the comparison group, the level of school, and variables related to instruction. Other variables we considered including in the model were *study quality*, *number of sessions* as a proxy for length of treatment, and *graphic organizers*. However, we elected not to include *study quality* because the correlation with effect size was almost zero. We also elected not to include *number of sessions* because we had to estimate this variable for a large number of studies (k = 8). Finally, we elected not to include *graphic organizers* because they were included in an overwhelming majority of the studies (75%) and had a small and nonsignificant correlation with the effect sizes (r = .05).

Results of the metaregression are provided in Table 6. *Number of structures taught*, *writing*, and *competing treatments* were significant contributors to the model with alpha set to .05. The model indicated an expected increase in the effect size of 0.13 for each additional text structure taught (in addition to the first one). Studies that included a writing component resulted in significantly larger effect sizes (B = 0.38) than studies that did not include a writing component. The large effect for writing illustrates the importance of controlling for additional moderators in a metaregression, as the correlation between writing and effect size was nonsignificant and only .10, but had a larger coefficient than the variable for the

Table 4
Bivariate Correlation Matrix of Study Characteristics and Effect Sizes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. ES (g)	1.00													
2. Publication type	.03	1.00												
3. Experiment	−.004	.46*	1.00											
4. Adjusted data	.08	−.25	−.68*	1.00										
5. Quality	−.001	.48*	.59*	−.19	1.00									
6. School level	−.02	−.003	−.16	−.09	−.11	1.00								
7. LD or at-risk	.23	−.22	−.02	−.01	.04	.09	1.00							
8. Explicit instruction	−.01	.04	−.07	.14	.23	−.14	.06	1.00						
9. # TSs taught	.24	−.56*	−.27	−.01	−.52*	−.02	.27	−.08	1.00					
10. # Sessions	.04	−.12	−.12	.11	−.12	−.17	−.24	.33*	.07	1.00				
11. Signal words	.01	.25	.01	.12	−.02	−.25	−.11	.28	.05	.22	1.00			
12. Graphic organizers	.05	.03	.06	.16	.23	−.03	−.34*	−.06	−.27	.24	−.06	1.00		
13. Writing	.10	−.14	−.23	.37*	.01	−.20	−.01	.37*	−.13	.37*	.23	.03	1.00	
14. Passage source	.15	.09	.37*	−.19	.17	−.19	−.14	−.21	−.22	−.24	.04	.09	−.19	1.00

Note. k = 40. ES = effect size; LD = learning disability; TSs = text structures.
* Statistically significant (p < .05).

Table 5
Average Weighted Effect Sizes and Confidence Intervals for TSI on Proximal Reading Outcomes (RQ1a)

Type of analysis	*k*	*N*	ES	95% CI	*z* score	*p* value	*Q* value	I^2	τ^2
Robust variance estimation	40	170	.57	[.39, .76]		<.001			.15
Aggregated ES analysis	40	40	.56	[.43, .69]	8.32	<.001	118.60	67.12	.07

Note. TSI = text structure instruction; RQ = Research Question 1a; *k* = number of studies; ES = effect size; CI = confidence interval.

number of structures taught, which had a larger direct correlation with effect size (r = .24).

The variable for *competing treatments*, on the other hand, was significant but negative. This indicated that comparing TSI with a competing treatment designed to improve reading comprehension may result in a significant reduction in the effect size. This should not be interpreted to mean that the effect size is expected to be null or negative when TSI is compared with competing comprehension treatments. As we might expect, however, the effect size is expected to be larger when comparing TSI with BAU.

In some cases, the null results of a metaregression are as important as the significant results. In this model, the coefficients for *school level, explicit instruction,* and *signal words* were all nonsignificant. For *school level*, the nonsignificant coefficient indicates that TSI is equally effective across both elementary and secondary grade levels. For *explicit instruction* and *signal words,* the nonsignificant effects should not be interpreted as indicating these instructional tools are not important. Rather, using signal words and explicit instruction techniques in TSI instruction is not expected to produce larger effects than interventions that use other instructional approaches used in these studies.

RQ1c: Analysis of studies that included only students with or at-risk for learning disabilities. Only eight studies included only students with or at-risk for learning disabilities or provided information that allowed us to calculate an effect size for these students. We elected not to include this variable in the metaregression because doing so would not provide a direct comparison of the effects of TSI for students with and without disabilities. However, students with learning disabilities are a particularly vulnerable population, and time should be spent on the most effective interventions for these students. Therefore, we felt it was necessary to include a separate analysis of these eight studies to determine whether TSI was effective with this population.

We aggregated the effects across subgroups and measures for this analysis. Seven of eight studies resulted in positive effects for this subgroup. Results of the model are presented in Table 7. The random effect analysis resulted in a significant average weighted effect size of 0.96, 95% CI [0.44, 1.47], indicating TSI has a large effect for this population of students. The test for heterogeneity indicated a *Q*-value of 21.64, which was larger than the critical value of 12.59 for a chi-square distribution with 7 degrees of freedom at $\alpha = .05$. The I^2 for the sample indicates that 67% of the variance was due to between study factors. Therefore, we rejected the null hypothesis of homogeneity within the sample. There were not enough studies to analyze between-study variance statistically. However, inspection of these studies revealed the two studies with the smallest effect sizes (McLaughlin, 1990; Wilkins, 2007), one of which was negative, showed only one text structure was taught in both studies, while all other studies included instruction in at-least two text structures. Furthermore, TSI was compared with a competing treatment in both of those studies, as well as the study with the third smallest effect size (Bakken et al., 1997, Study 2). Given the results of the metaregression results for RQ1b, it is seems likely that these two variables may explain some of the variance between studies included for this research question.

Summary of the Results for Research Question 2: Effects of TSI on Transfer Outcomes

As discussed in the introduction, we framed transfer in terms of the taxonomy of transfer described by Barnett and Ceci (2002). Based on the studies identified for inclusion in this meta-analysis, we were able to classify studies in relation to three types of transfer. First, we examined transfer of skills in *temporal context* (i.e., maintenance of the effects over time; RQ2a). Second, we examined near-transfer to the *knowledge domain* (i.e., understanding of one expository text

Table 6
Metaregression for RQ1 Using Robust Variance Estimation and Small Sample Corrections

Variable	*B*	*SE*	*df*	*p*	95% CI
Intercept	.24	1.16	15.36	.156	[−.10, .58]
Competing treatment (1 = yes)	−.39	.17	13.66	.036	[−.75, −.03]
School level (secondary = 1)	.07	.13	17.68	.584	[−.21, .35]
Structures taught (1–5)[a]	.13	.05	10.69	.025	[.02, .25]
Explicit instruction (yes = 1)	−.24	.14	11.98	.105	[−.54, .06]
Writing (1 = yes)	.38	.12	10.65	.010	[.11, .65]
Signal words (1 = yes)	.09	.14	16.11	.542	[−.21, .38]

Note. RQ1 = Research Question 1; CI = confidence interval. The model is based on k = 40 studies and N = 170 effect sizes. τ^2 = .158.
[a] The structures taught represents five text structures, but was actually coded as 0–4 to aid in the interpretability of the coefficient, as it was not possible for a study in this analysis to teach fewer than one text structure.

Table 7
Average Weighted Effect Sizes for Studies Involving Only Students With or At-Risk for Learning Disabilities

				Test of null hypothesis		Heterogeneity	
Research question	*k*	ES	95% CI	*z* score	*p* value	*Q* value	I^2
RQ1c	8	.96	[.44, 1.47]	3.61	<.001	21.64	67.65

Note. *k* = number of studies; CI = confidence interval; RQ1c = Research Question 1c.

structure influencing comprehension in expository text written in an untaught structure; RQ2b). Third, we examined far-transfer across the *knowledge domain* and *modality* (i.e., to norm-referenced measures of general reading comprehension; RQ2c).

In the cases of RQ2b and RQ2c, we could not use robust variance estimation due to having less than 10 studies in the analyses (see Tanner-Smith & Tipton, 2014). For the sake of consistency, we used aggregated effect sizes across groups and measures for all of the RQ2 questions. The results for the analyses related to RQ2a, RQ2b, and RQ2c are presented in Table 8.

RQ2a (temporal transfer): Maintenance of the effects over time. Eleven studies met the criterion for inclusion in RQ2a. Delayed posttests were administered one day to three months following the posttest. A majority of the studies ($k = 9$) included delayed posttests administered 1 week or longer following the posttest.

The analysis resulted in a significant average weighted effect size of 0.57, 95% CI [0.26, 0.87]. Nine of the 11 studies included in this analysis resulted in positive effects for TSI. This evidence indicates that the effects for TSI are maintained over period time; however, the median time period between the posttest and the delayed posttest was 7 days, so we cannot make strong inferences about longer periods of time. The test for heterogeneity resulted in a *Q*-value of 15.97, which is smaller than the critical value of 19.68 for a model with 11 degrees of freedom, indicating that this level of variation between studies could be expected by chance. Therefore, we could not reject the null hypothesis of homogeneity within the sample of studies.

We conducted a sensitivity analysis of these findings by creating new effect sizes subtracting the posttest outcomes from the delayed posttest outcomes. We then calculated an average weighted effect size. The analysis resulted in a positive significant average weighted effect size of 0.22, 95% CI [0.04, 0.41], indicating that the results are being maintained.

RQ2b (near-transfer within the knowledge domain): Transfer of TSI skills to comprehension of expository text using an untaught structure. Seven studies met the additional criterion for inclusion in RQ2b. The structures taught and the untaught structure assessed varied widely across these studies. In five of the seven studies, TSI involved instruction in a single structure and students' comprehension was assessed in a second structure, whereas in two of the studies, TSI involved instruction in two structures and students' comprehension was assessed in a third structure. The structures taught across the studies include *description, sequence, cause/effect,* and *compare/contrast.* The structures assessed include *sequence, description, and cause/effect*, as well as a structure one of the authors identified as "unstructured text."

The analysis resulted in a nonsignificant average weighted effect size of 0.62, 95% CI [0.01, 1.23]. Only four of seven studies (57%) relevant to answering this question resulted in positive effects. Although the tests for heterogeneity resulted in a *Q*-value of 29.05, which is larger than the critical value of 12.59 for a model with 6 degrees of freedom, this comparison involved only seven studies, not meeting our criteria for follow-up moderator analyses. However, the large confidence interval confirmed the large amount of variation between studies included in this analysis.

RQ2c (far-transfer within the knowledge domain and across modalities): Transfer of TSI skills to norm-referenced measures of general reading comprehension. We classified the use of norm-referenced, general reading comprehension measures as far-transfer. Students were asked to transfer knowledge across the knowledge domain (i.e., general reading as opposed to expository reading) and functional contexts (i.e., norm-referenced tests as opposed to classroom work). Eight studies met the additional criterion for inclusion in RQ2c. The analysis resulted in a significant average weighted effect size of 0.13, 95% CI [.03, 0.25]. However, only four of the eight studies (50%) resulted in positive effect sizes, giving us less confidence in this finding. For this analysis, the test for heterogeneity indicated a *Q*-value of 13.83, which is smaller than the critical value of 15.51 for a chi-square distribution with 7 degrees of freedom. We could not reject the null hypothesis of homogeneity within the sample.

Table 8
Average Weighted Effect Sizes for Transfer Analyses (RQ2)

				Test of null hypothesis		Heterogeneity	
Research question (RQ)	*k*	ES	95% CI	*z* score	*p* value	*Q* value	I^2
RQ2a: Temporal transfer (maintenance)	11	.57	[.26, .87]	3.65	<.001	15.87	39.99
RQ2b: Near-transfer untaught structure	7	.62	[.007, 1.23]	1.981	.048	29.05	79.34
RQ2c: Far transfer general reading comprehension	8	.13	[.03, .25]	2.58	.010	13.83	49.42

Note. *k* = number of studies; ES = effect size; CI = confidence interval.

Publication Bias Analyses

We conducted a fairly extensive search of the literature and a majority of studies we included (53%, $k = 24$) came from dissertations, technical reports, and conference papers. However, some unpublished studies with small effects may not have been located. The result of the our four publication bias analyses gave mixed results (see Appendix C), indicating that there may be some unpublished studies with null or negative effects missing from our analysis. If so, our overall average weighted effect size for RQ1a might be slightly overestimated. However, this is not likely to impact the interpretation of the results.

Discussion

Reading is one of the primary ways people are introduced to new information. Authors of informational text often use devices to present and organize information in ways that will help their reader understand it better. Text structure is one such device (Kintsch, 1974; Meyer, 1975). Knowing how authors decided to structure a text may provide readers with valuable information about how to approach the text and assist them in identifying important information to remember from the text. Meyer (1985) identified five text structures, including description, compare/contrast, sequence, cause/effect, and problem/solution. Researchers have argued: (a) knowledge of text structures may be beneficial for improving expository reading skills for all students, including students in elementary and secondary schools and students with disabilities; (b) specific instructional factors may play a role in the effectiveness of text structure instruction (e.g., grade level, number of text structures taught, signal words); and (c) TSI may lead to transfer of skills.

There were two purposes for conducting this meta-analysis. One, to determine whether TSI is effective for improving proximal measures of informational text comprehension and examine factors that might moderate the effects of TSI. Two, to determine whether the effects of TSI transfer temporally, to near-transfer measures of reading comprehension of expository text written in an untaught structure, and/or to far-transfer measures of general reading comprehension.

Widely cited standards for assessing the magnitude of effect sizes in behavioral and social science are the values presented by Cohen (1977) indicating that an effect size of .20 is "small," .50 is "medium," and .80 is "large." However, Lipsey et al. (2012) suggested that representing effect sizes in such a way can be inappropriate and/or misleading. They provided a guide for representing educational effects in more meaningful ways. To that end, we attempted to characterize the effect sizes presented in this review in relation to the distribution of other mean effect sizes in the same general area (i.e., other treatments designed to influence reading performance), and in terms of performance on specific measures.

TSI Improves Proximal Measures of Expository Reading Comprehension

The evidence from this meta-analysis indicates that teaching students about expository text structures improves their expository reading comprehension across all comparison groups used in the studies located for this review. The significant overall average weighted effect size for TSI using robust variance estimation on researcher created measures of reading comprehension was 0.57, which is larger than the average effect size found across educational interventions examining effects on researcher developed measures ($ES = 0.39$; Lipsey et al., 2012).

Confidence can be placed in the findings for RQ1a, as positive effects of the TSI were replicated repeatedly in 85% of the studies. However, the quality of the studies was relatively moderate, tempering the findings to a degree. Overall, the studies received an average quality score .57 for RQ1a. The primary weaknesses across the studies were the failure to report fidelity and failure to include more than one teacher per condition. It is difficult to determine if the first issue is critical, as a lack of fidelity reporting does not mean the interventions were implemented poorly, or even that the researchers did not collect fidelity data, simply that they did not report it. The second issue should be less of a concern across a large body of studies. Despite both concerns, the consistency of the findings lends support for the theory that instruction in expository text structures can improve expository reading comprehension.

An important finding to note from the metaregression analyses was that instruction in more text structures resulted in statistically significantly larger effects. The coefficient indicated an expected 0.13 standard deviation increase for each text structure taught after the first one. It is important to remember that the variable representing instruction in text structures was not a proxy for the length of the study, as the correlation between *number of structures taught* and *number of sessions* was .07. A qualitative examination of the types of text structures taught, the order in which they were taught, and specific groupings of text structures taught revealed that there was quite a bit of variation in the approach to teaching multiple text structures. However, it is interesting to note that when only one text structure was taught, it was most likely to be *compare/contrast* (46%) or *cause/effect* (23%). These two text structures were also more likely to be included in the instruction when only two text structures were taught (*compare/contrast* = 73%, *cause/effect* = 20%). The percentages for *compare/contrast* are particularly striking, and may confound the analysis between the number of structures taught and the type of structures taught. That is, it may be that students are more familiar with the compare/contrast text structure than other structures, as the concept of comparing and contrasting is often taught across subject areas (e.g., scientific classification systems, mathematics equations, narrative and expository reading instruction using Venn diagrams). If so, teaching this expository text structure to students may result in smaller effects, while teaching them text structures they are less familiar with may lead to larger gains. Thus, teaching more structures in addition to *compare/contrast* would also lead to larger gains. That said, it seems logical that knowledge in more text structures would improve comprehension over knowledge of a single structure.

Another important finding of the metaregression model was that *writing* was a significant predictor of the effectiveness of the intervention. For the sake of this study, writing was considered to include note-taking, sentence writing (e.g., when answering short answer questions), and writing paragraph-length responses to text. Studies that included writing were expected to increase the effect size by an average of 0.38 standard deviation units. These results are consistent with recent findings that writing is effective for improving reading outcomes (Graham & Hebert, 2010). It should be noted that this finding does not indicate that studies without writing were not effective (75% of studies that did not include writing resulted in positive effect sizes). Rather, it seems that writing simply enhanced text

structure instruction across studies when controlling for the other variables in the metaregression model.

A third significant predictor in the metaregression model was the variable for *competing treatments*. The negative coefficient and confidence interval indicated that the effect size is expected to be significantly smaller when TSI is compared with a competing comprehension treatment. This is not exactly surprising, as we expect smaller effect sizes when interventions are compared with other effective approaches than when they are compared with BAU. However, illustrates that the effectiveness of TSI may be tempered by the comparison to a competing treatment and may have implications for instructional choices.

In addition to the statistically significant findings of the metaregression, nonsignificant findings of the moderator analysis provided evidence that the effects of TSI are robust across levels of schooling and the instructional variables for *explicit instruction* and *signal words*. It is important to note that these nonsignificant findings do not indicate that these variables do not influence the effects of TSI in important ways, but that TSI is just as effective when these variables vary across studies.

Finally, the breakout analysis was conducted for studies involving students with or at-risk for disabilities indicated that TSI was effective for this population of students. Seven of eight studies resulted in positive effects for these students, showing fairly consistent evidence for this finding. This is important, because instructional time needs to be used wisely for this population. More research needs to be conducted in order to understand how different combinations of instructional approaches can be applied to maximize the effects for these students. Although there were not enough studies to examine the effects of potential moderators, examination of the coding showed that *number of structures taught* and *competing treatments* may explain some of the variability between studies, similar to the moderator analysis conducted in the metaregression for RQ1b.

The Effects of TSI Transfer Across Three Contexts: Temporal, Near-Knowledge Domains, and Far-Knowledge Domains and Modalities

Evidence from the analyses of the three RQ2 subquestions indicated that text structure knowledge transfers to other contexts. First, analysis of effect sizes related to delayed posttest measures indicated that the effects of TSI transfer across temporal contexts. However, this finding must be interpreted cautiously, as the median delay between the posttests and maintenance measures across studies was only seven days. We cannot assume that the effects of TSI can be maintained over longer periods of time.

Positive effects were also found for TSI on near-transfer measures of the knowledge domain. The effect size on measures of transfer from comprehension of taught structures to untaught structures was 0.62. One possible explanation for the transfer results might be that some structures have similar features, resulting in transfer effects due to similarities between the structures taught and assessed. For example, authors using a *cause/effect* text structure to provide information about a phenomenon may be likely to do this temporally, indicating the cause(s) first, followed by the effect(s). Using such an approach is similar to the *sequence* text structure. Thus, students learning about one text structure may be more likely to improve in a second text structure. Another explanation is that instruction in one text structure may bring students' attention to features and organization of other structures more generally.

Finally, the analysis showed that TSI resulted in far-transfer to norm-referenced measures of reading (i.e., far-transfer across the knowledge domain and across modalities from the informal instructional tasks to formal testing tasks). Despite the significant ES, however, we have less confidence in this finding because 44% of the studies (four of nine) resulted in negative effects.

The average weighted effect size for norm-referenced measures of reading comprehension was 0.15. For comparison sake, Lipsey et al. (2012) reported an average effect size of 0.24 obtained across educational interventions examining effects on standardized measures analogous to the norm-references measures of general reading comprehension used for this analysis. For more specific context, however, the impact of TSI is similar to the effect size of 0.17 found for reading programs in middle and high school (Slavin, Cheung, Groff, & Lake, 2008) and the effect size of 0.10 found for vocabulary instruction (Elleman, Lindo, Morphy, & Compton, 2009). However, the effect for TSI is smaller than the effect size of 0.37 found for writing (Graham & Hebert, 2010, 2011), and the effect size of 0.32 for reciprocal teaching (Rosenshine & Meister, 1994). This lends support for our previous consideration that the effectiveness of TSI likely depends on factors such the strength of the instructional condition to which it is compared.

Implications and Recommendations for Research and Practice

The consistency and magnitude of the effects of TSI on researcher created expository reading outcomes and transfer measures were notable in this review. We have four recommendations based on our analysis.

One, we recommend that text structure instruction be included as one component of a comprehensive approach to expository reading instruction. The analyses indicated TSI is effective across elementary and secondary grade levels, and compared with all of the comparison conditions found in the studies. Given the larger number of effective teaching practices for improving reading comprehension, and the potential of the effects of TSI to vary depending on the instruction to which it is compared.

Two, teachers should provide instruction in multiple text structures. Teaching more text structures resulted in larger effect sizes. Although the results also indicated that instruction in one text structure improves comprehension in an untaught text structure, the results do not suggest whether this is true for all structures or whether there is a strong relationship between particular structures. Our observations of the structures taught revealed that researchers studied the effectiveness of compare/contrast structure most often when studying a single text structure or only two text structures. It may be that the *compare/contrast* structure is chosen in this context because it is viewed as easier to teach or learn. Further research needs to be conducted to determine whether some text structures are easier to learn than others, whether some combinations of text structures are more effective than others, whether some text structures complement one another for instructional purposes, or whether there is an optimal order to teach the structures.

Three, it is important for teachers and researchers to include writing as a part of the instructional approach for text structure

instruction. Writing has been shown to be effective for improving reading outcomes (see Graham & Hebert, 2010, 2011), and it was a particularly strong moderator of the effect for TSI. The studies in this review included writing such as note taking, sentence length writing (such as writing answers to questions), and paragraph-length writing (such as writing summaries of text). More research needs to be conducted on how to maximize the benefits of combining writing with TSI, but in the meantime, the strength of these effects should not be overlooked.

Finally, more research needs to be conducted on instructional factors such as explicit instruction and signal words. *Signal words* was not a significant moderator of the ES, nor was *explicit instruction*. However, the moderator analyses conducted in this review were broad and limited by the available literature. The power to examine moderators in the metaregression model was also limited (by the number of studies, not the number of effect sizes) so we could not include interaction variables such as the interaction between instructional variables and grade level. It may be that signal words are important for younger students, but not older students. For example, Kao and Williams (2015) stated that controlling text when teaching younger students text structures might be analogous to using nonsense words in decoding instruction, but it is unclear whether this is true for older readers. Similarly, use of explicit instruction may be effective for some populations, but not others. Further research needs to be conducted to understand such potential nuances.

Caveats and Limitations

As with any meta-analysis, the choices we made for inclusion and exclusion of studies, coding procedures, and moderator analyses, among other factors, limit our findings in important ways. First, this meta-analysis was designed to draw conclusions about our research questions. The generalizability of such conclusions is limited based on a variety of factors such as the participants in the studies, quality of the investigations, outcome measures used, and so on. For example, additional research is also needed to help us determine whether and how characteristics of instruction, such as use of graphic organizers or writing, may interact with participant characteristics such as grade level or reading ability.

Second, a concern with meta-analysis involves the comparability of the outcome measures on which the effect sizes are based. To contend with this, we analyzed only expository reading comprehension outcome measures and norm-referenced measures of general reading comprehension, eliminating measures for other constructs (e.g., decoding, reading fluency, writing). We also analyzed our outcome measures on a continuum of transfer from proximal measures of comprehension to near-transfer temporal measures, near-transfer knowledge domain measures, and far-transfer knowledge domain measures. Other researchers may have chosen different sets of measures.

Third, there was quite a bit of variability in the text structure treatments, some of which we were able to examine, and others that we did not have the power to examine or were not reported consistently by authors. For example, the number of structures taught, use of signal words, use of graphic organizers, inclusion of writing, the text used in the studies, and the content areas, among other factors, are all likely to play a role in the effectiveness of the instruction. Although we were able to analyze some sources of potential variability within the studies, we simply did not have enough studies to explore all of these potential moderators.

Finally, we must note the limited external validity of the research on text structure instruction. Eighty percent of the studies in this review involved fewer than 100 subjects. Effects found for such small studies may not generalize to examinations of the effectiveness of text structure instruction conducted in larger and more diverse samples. Moreover, of the nine studies that involved sample sizes of larger than 100, one research lab conducted four of those studies (Williams et al., 2005; Williams et al., 2007; Williams et al., 2014; Williams et al., 2009) and another research lab conducted two of those studies (Wijekumar et al., 2012; Wijekumar et al., 2014). To highlight the issue of generalizability, the studies conducted by Williams and colleagues were all conducted in second grade classrooms, while the studies conducted by Wijekumar and colleagues were computer-based interventions that are likely to be practically different than teacher led interventions. That said, all nine of the studies with larger samples resulted in positive effects.

Future Research

This review provided important insight into the strengths and weaknesses of the experimental literature examining the effectiveness of text structure instruction for improving students' reading comprehension. Although a considerable number of studies have been conducted there are several gaps in the research base. As mentioned before, more research is needed focusing on the instructional aspects of text structure interventions and whether certain aspects of instruction in text structures are differentially effective for specific subgroups of students or age ranges.

Also, in a more subjective look at the studies that included instruction in multiple text structures, it appears that the structures were taught in a serial fashion, but there does not seem to be a convention for order of text structures taught. Future research in this area should attempt to determine whether students benefit from instruction in these structures in a particular order (does instruction in one text structure facilitate learning of another specific structure?) or even whether teaching the structures serially or concurrently is more effective.

Finally, although there were many well implemented research studies (especially more recently), a large majority of investigators failed to report fidelity of implementation, less than one half of the studies included randomization with analysis at the correct level, more than one half of the studies included only one teacher per condition, and some studies omitted specifics on the demographic characteristics of their student samples. Additional high quality studies need to be conducted to increase confidence in the generalizability of the findings. To address another limitation of this literature, future studies should be conducted with large and diverse samples.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Alvermann, D. E. (1981). The compensatory effect of graphic organizers on descriptive text. *The Journal of Educational Research, 75,* 45–48. http://dx.doi.org/10.1080/00220671.1981.10885354

*Alvermann, D. E. (1982). Restructuring text facilitates written recall of main Ideas. *Journal of Reading, 25,* 754–758.

*Alvermann, D. E., & Boothby, P. R. (1984, April). *Knowledge of text structure and its influence on a transfer task*. Paper presented at the 68th Annual Meeting of the American Education Research Association. New Orleans, LA.

Armbruster, B. B., & Anderson, T. H. (1980). *The effect of mapping on the free recall of expository text* (Tech. Rep. No. No. 160). Cambridge, MA: Bolt, Beranek and Newman.

*Bakken, J. P., Mastropieri, M. A., & Scruggs, T. E. (1997). Reading comprehension of expository science material and students with learning disabilities: A comparison of strategies. *The Journal of Special Education, 31,* 300–324. http://dx.doi.org/10.1177/002246699703100302

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based Writing-to-Learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74,* 29–58. http://dx.doi.org/10.3102/00346543074001029

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128,* 612–637.

*Bartlett, B. J. (1978). *Top-level structure as an organizational strategy for recall of classroom text* (Unpublished dissertation). Arizona State University, Phoenix, AZ.

*Bartlett, B. J., Turner, A., & Mathams, R. (1980). *Top-level structure: A significant relation for what fifth-graders remember from classroom reading*. (ERIC Document Reproduction Service No. ED200932).

Bohaty, J. J. (2015). *The effects of a standard protocol intervention on the reading outcomes of 4th and 5th graders experiencing reading difficulties* (Unpublished dissertation). The University of Nebraska–Lincoln, Lincoln, NE.

Bohaty, J., Hebert, M., Nelson, J. R., & Brown, J. A. (in press). Methodological status and trends in expository text structure instruction efficacy research. *Reading Horizons*.

Brandt, D. M. (1978). *Prior knowledge of the author's schema and the comprehension of prose* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations. (AAT 7911123)

*Brimmer, K. M. (2004). *Using thinking aloud procedures to promote reading comprehension of expository texts with intermediate grade level students* (Unpublished dissertation). Oakland University, Rochester, MI.

*Broer, N. A., Aarnoutse, C. A. J., Kieviet, F. K., & van Leeuwe, J. F. J. (2002). The effects of instructing the structural aspects of text. *Educational Studies, 28,* 213–238. http://dx.doi.org/10.1080/0305569022000003681

Ciullo, S., Lo, Y. L. S., Wanzek, J., & Reed, D. K. (2016). A synthesis of research on informational text reading interventions for elementary students with learning disabilities. *Journal of Learning Disabilities, 49,* 257–271.

Cohen, J. (1977). *Statistical power analyses for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.

*Coleman, P. A. B. (1983). *Effects of graphic organizers, text organization, and reading ability on the recall of text information* (Unpublished dissertation). The University of Michigan, Ann Arbor, MI.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs* (Vol. 129). Thousand Oaks, CA: Sage.

*Crowe, E., Al Otaiba, S., & Lonnigan, C. J. (2014, March). *The rise and fall (and rise again) of a small-group instructional program to teach students text structures to promote comprehension: Results of two randomized studies*. Paper presented at the Society for Research on Educational Effectiveness Conference. Washington, DC.

*Duffy, J. (1985). *Effects of instruction in noting and using an author's pattern of writing on sixth graders' understanding and memory of expository text*. (Unpublished Dissertation). Temple University.

Duke, N. K. (2000). 3.6 minutes per day: The scarcity of informational texts in first grade. *Reading Research Quarterly, 35,* 202–224. http://dx.doi.org/10.1598/RRQ.35.2.1

Duke, N., Pearson, D., Strachan, S., & Billman, A. (2011). Essential elements of fostering and teaching reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 51–93). Newark, DE: International Reading Association.

Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 205–242). Newark, DE: International Reading Association.

Duke, N. K., & Roberts, K. L. (2010). The genre-specific nature of reading comprehension and the case of informational text. In D. Wyse, R. Andrews, & J. Hoffman (Eds.), *The international handbook of English language and literacy teaching* (pp. 74–86). London, UK: Routledge.

Elleman, A., Lindo, E., Morphy, P., & Compton, D. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness, 2,* 1–44. http://dx.doi.org/10.1080/19345740802539200

Englert, C. S., & Hiebert, E. H. (1984). Children's developing awareness of text structures in expository materials. *Journal of Educational Psychology, 76,* 65–74. http://dx.doi.org/10.1037/0022-0663.76.1.65

*Englert, C. S., Raphael, T. E., Anderson, L. M., Anthony, H. M., & Stevens, D. D. (1991). Making strategies and self-talk visible: Writing instruction in regular and special education classrooms. *American Educational Research Journal, 28,* 337–372. http://dx.doi.org/10.3102/00028312028002337

Englert, C. S., & Thomas, C. C. (1987). Sensitivity to text structure in reading and writing: A comparison between learning disabled and non-learning disabled students. *Learning Disability Quarterly, 10,* 93–105. http://dx.doi.org/10.2307/1510216

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

Garner, R., Alexander, P., Slater, W., Hare, V. C., Smith, T., & Reis, R. (1986). Children's knowledge of structural properties of expository text. *Journal of Educational Psychology, 78,* 411–416. http://dx.doi.org/10.1037/0022-0663.78.6.411

*Gentry, L. J. (2006). *Comparison of the effects of training in expository text structure through annotation textmarking and training in vocabulary development on reading comprehension of students going into fourth grade* (Unpublished dissertation). University of South Florida, Tampa, FL.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71,* 279–320. http://dx.doi.org/10.3102/00346543071002279

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York, NY: Russell Sage Foundation.

*Gould, B. T. (1987). *Effects of prior knowledge and text structure instruction on the comprehension and memory for expository reading of intermediate and junior-high grade students* (Unpublished dissertation). Boston University, Boston, MA.

Graham, S., & Hebert, M. (2010). *Writing to read: The evidence-base for how writing can improve reading*. Washington, DC: Alliance for Excellent Education (Manuscript commissioned by the Carnegie Corporation of New York).

Graham, S., & Hebert, M. (2011). Writing-to-read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review, 81,* 710–744. http://dx.doi.org/10.17763/haer.81.4.t2k0m13756113566

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104,* 879–896. http://dx.doi.org/10.1037/a0029185

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high school. A report to*

Carnegie Corporation of New York. Washington, DC: Alliance for Excellent Education.

Griffin, C. C., & Tulbert, B. L. (1995). The effect of graphic organizers on students'comprehension and recall of expository text: A review of the research and implications for practice. *Reading & Writing Quarterly, 11,* 73–89. http://dx.doi.org/10.1080/1057356950110106

*Hall, K. M. (2002). *A study of the effect of text structure and content on at-risk second graders' comprehension of compare/contrast expository text* (Unpublished dissertation). Columbia University, New York, NY.

*Hall, K. M., Sabey, B. L., & McClellan, M. (2005). Expository text comprehension: Helping primary-grade teachers use expository texts to full advantage. *Reading Psychology, 26,* 211–234. http://dx.doi.org/10.1080/02702710590962550

*Hammann, L. A. (2000). *An investigation of instruction in summarizing and text structure for compare and contrast writing* (Unpublished dissertation). The Pennsylvania State University, State College, PA.

*Hammann, L. A., & Stevens, R. J. (2003). Instructional approaches to improving students' writing of compare-contrast essays: An experimental study. *Journal of Literacy Research, 35,* 731–756. http://dx.doi.org/10.1207/s15548430jlr3502_3

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 37–48). New York, NY: Russell Sage Foundation.

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. http://dx.doi.org/10.1002/jrsm.5

*Hickerson, B. L. (1986). *Critical thinking, reading and writing: Developing a schema for expository text through direct instruction in analysis of text structure* (Unpublished dissertation). North Texas State University, Denton, TX.

*Hoffman, K. F. (2010). *The impact of graphic organizer and metacognitive monitoring instruction on expository science text comprehension in fifth grade students* (Unpublished dissertation). North Carolina State University, Raleigh, NC.

Kao, J., & Williams, J. (2015, February). *Improving instruction in reading and writing: Critical factors and acute challenges*. Paper presented at the Pacific Coast Research Conference, San Diego.

Kieras, D. (1978). Beyond pictures and words: Alternative information-processing models for imagery effects in verbal memory. *Psychological Bulletin, 85,* 532–554. http://dx.doi.org/10.1037/0033-2909.85.3.532

Kim, A. H., Vaughn, S., Wanzek, J., & Wei, S. (2004). Graphic organizers and their effects on the reading comprehension of students with LD: A synthesis of research. *Journal of Learning Disabilities, 37,* 105–118. http://dx.doi.org/10.1177/00222194040370020201

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

Lapp, D., Flood, J., & Ranck-Buhr, W. (1995). Using multiple text formats to explore scientific phenomena in middle school classrooms. *Reading & Writing Quarterly, 11,* 173–186. http://dx.doi.org/10.1080/1057356950110206

Lipsey, M., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. (NCSER 2013–3000). Washington, DC: U.S. Government Printing Office.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9,* 111–151. http://dx.doi.org/10.1016/0010-0285(77)90006-8

*McDermott, M. B. (1990). *Teaching top level structure as an aid for reading comprehension of expository prose for fourth grade pupils* (Unpublished dissertation). Hofstra University, New York, NY.

*McLaughlin, E. M. (1990). *Effects of graphic organizers and levels of text difficulty on less-proficient fifth-grade readers' comprehension of expository text* (Unpublished dissertation). University of Maryland, College Park, MD.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam, Netherlands: North-Holland Publishing.

Meyer, B. J. F. (1979). Organizational patterns in prose and their use in reading. In M. L. Kamil & A. J. Moe (Eds.), *Reading research: Studies and applications* (pp. 109–117). National Reading Conference: Clemson, SC.

Meyer, B. J. F. (1985). Prose analysis: Purposes, procedures, and problems. In B. K. Britton & J. Black (Eds.), *Understanding expository text: A theoretical and practical handbook for analyzing explanatory text* (pp. 269–304). Hillsdale, NJ: Erlbaum.

Meyer, B. J. F. (1987). Following the author's top-level organization: An important skill for reading comprehension. In R. J. Tierney, P. L. Anders, & J. Nichols Mitchell (Eds.), *Understanding readers' understanding: Theory and practice* (pp. 59–76). Hillsdale, NJ: Erlbaum.

Meyer, B. J. F., Brandt, D. M., & Bluth, G. J. (1980). Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly, 16,* 72–103. http://dx.doi.org/10.2307/747349

*Meyer, B. J. F., Middlemiss, W., Theodorou, E., Brezinski, K. L., McDougall, J., & Bartlett, B. J. (2002). Effects of text structure strategy instruction delivered to fifth-grade children using the internet with and without the aid of older adult tutors. *Journal of Educational Psychology, 94,* 486–519. http://dx.doi.org/10.1037/0022-0663.94.3.486

Meyer, B. J. F., & Ray, M. N. (2011). Structure strategy interventions: Increasing reading comprehension of expository text. *International Electronic Journal of Elementary Education, 4,* 127–152.

Meyer, B. J., & Rice, G. E. (1984). The structure of text. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 319–351). New York, NY: Longman, Inc.

*Moore, S. R. (1995). *Building schemata for expository text through collaboration and an integration of reading and writing* (Unpublished dissertation). Harvard University, Cambridge, MA.

*Moore, S. R. (1996, April). *Collaboration and the reading-writing relationship: Implications for building schemata for expository text*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

*Newman, L. M. (2007). *The effects of explicit instruction of expository text structure incorporating graphic organizers on the comprehension of third-grade students* (Unpublished dissertation). Eastern Washington University, Cheney, WA.

Nouri, H., & Greenberg, R. H. (1995). Meta-analytic procedures for estimation of effect sizes in experiments using complex analysis of variance. *Journal of Management, 21,* 801–812. http://dx.doi.org/10.1177/014920639502100411

*Occasio, T. L. (2006). *A comparison of two instructional programs to develop strategies to improve reading comprehension* (Unpublished dissertation). Widener University, Chester, PA.

Ordynans, J. G. (2012). *The effectiveness of inserted strategy questions on elementary students' comprehension of well-structured and less-structured expository text* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations. (AAT 3502558)

Piasta, S. B., & Wagner, R. K. (2010). Developing Early Literacy Skills: A Meta-Analysis of Alphabet Learning and Instruction. *Reading Research Quarterly, 45,* 8–38. http://dx.doi.org/10.1598/RRQ.45.1.2

Raphael, T. E., Englert, C. S., & Kirschner, B. W. (1986). *The impact of text structure instruction and social context on students' comprehension and production of expository text. Research Series No. 177*. East Lansing, MI: The Institute for Research on Teaching, Michigan State University.

*Reese, D. J. (1988). *Effect of training in expository text structure in reading comprehension* (Unpublished dissertation). University of South Florida, Tampa, FL.

*Reynolds, G. A. (2006). *Teaching composing from sources to middle grade students* (Unpublished dissertation). Columbia University, New York, NY.

*Reynolds, G. A., & Perin, D. (2009). A comparison of text structure and self-regulated writing strategies for composing from sources by middle school students. *Reading Psychology, 30,* 265–300. http://dx.doi.org/10.1080/02702710802411547

Richgels, D., McGee, L. M., Lomax, R. G., & Sheard, C. (1987). Awareness of four text structures: Effects on recall of expository text. *Reading Research Quarterly, 22,* 177–196. http://dx.doi.org/10.2307/747664

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

*Russell, S. L. (2005). *Challenging task in appropriate text: Designing discourse communities to increase the literacy growth of adolescent struggling readers* (Unpublished dissertation). University of Maryland, College Park, MD.

Sáenz, L. M., & Fuchs, L. S. (2002). Examining the reading difficulty of secondary students with learning disabilities expository versus narrative text. *Remedial and Special Education, 23,* 31–41. http://dx.doi.org/10.1177/074193250202300105

*Samson, K. M. (1982). *An instructional strategy for helping readers identify the GIST in expository text* (Unpublished dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research, 84,* 328–364. http://dx.doi.org/10.3102/0034654313500826

*Scott, D. B. (2011). *Explicit instruction on rhetorical patterns and student-constructed graphic organizers: The impact on sixth-grade students' comprehension of social studies text* (Unpublished dissertation). University of Maryland, College Park, MD.

Shadish, W. R., Robinson, L., & Congxiao, L. (1999). *ES: A computer program for effect size calculation.* Memphis, TN: University of Memphis.

Siedow, M. D., & Fox, B. J. (1984). Effects of training on good and poor readers' use of top-level structure. *Literacy Research and Instruction, 23,* 340–346.

*Slater, W. H. (1985). Teaching expository text structure with structural organizers. *Journal of Reading, 28,* 712–718.

Slater, W. H. (1988). Current theory and research on what constitutes readable expository text. *Technical Writing Teacher, 15,* 195–206.

*Slater, W. H., Graves, M. F., & Piché, G. L. (1985). Effects of structural organizers on ninth-grade students' comprehension and recall of four patterns of expository text. *Reading Research Quarterly, 20,* 189–202. http://dx.doi.org/10.2307/747755

Slavin, R., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best evidence synthesis. *Reading Research Quarterly, 43,* 290–322. http://dx.doi.org/10.1598/RRQ.43.3.4

*Smith, P. L., & Friend, M. (1986). Training learning disabled adolescents in a strategy for using text structure to aid recall of instructional prose. *Learning Disabilities Research, 2,* 38–44.

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

*Spires, H. A., Gallini, J., & Riggsbee, J. (1992). Effects of schema-based and text structure-based cues on expository prose comprehension in fourth graders. *Journal of Experimental Education, 60,* 307–320. http://dx.doi.org/10.1080/00220973.1992.9943868

Swanson, H. L. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes.* New York, NY: Guilford Press.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5,* 13–30. http://dx.doi.org/10.1002/jrsm.1091

Taylor, B. M. (1982). Text structure and children's comprehension and memory for expository material. *Journal of Educational Psychology, 74,* 323–340. http://dx.doi.org/10.1037/0022-0663.74.3.323

*Taylor, B. (1985). Improving middle-grade students' reading and writing of expository text. *The Journal of Educational Research, 79,* 119–125. http://dx.doi.org/10.1080/00220671.1985.10885661

Taylor, B. M., & Beach, R. W. (1984). The effects of text structure instruction on middle-grade students' comprehension and production of expository text. *Reading Research Quarterly, 19,* 134–146. http://dx.doi.org/10.2307/747358

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20,* 375–393.

Ulper, H., & Akkok, E. A. (2010). The effect of using expository text structures as a strategy on summarization skills. In L. E. Kattington (Ed.), *Handbook of curriculum development* (pp. 303–328). New York, NY: Nova Science Publishers, Inc.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139,* 352–402. http://dx.doi.org/10.1037/a0028446

*Walker, M. L. (1991). *A study of the relative effectiveness of a textbook-study system (SQ3R), its variation (SRQ2R), and structure of text instruction* (Unpublished dissertation). Wayne State University, Detroit, MI.

Ward-Washington, R. (2001). *The effectiveness of instruction in using reading comprehension strategies with eleventh-grade social studies students* (Unpublished dissertation). Retrieved from ProQuest Digital Dissertations. (AAT 3040621)

What Works Clearinghouse. (2014). *WWC procedures and standards handbook: Version 3.0.* Washington, DC: Institute for Education Sciences.

*Whittaker, A. K. (1992). *Constructing science knowledge from exposition: The effects of text structure instruction* (Unpublished dissertation). Stanford University, Stanford, CA.

Wigent, C. A. (2013). High school readers: A profile of above average readers and readers with learning disabilities reading expository text. *Learning and Individual Differences, 25,* 134–140. http://dx.doi.org/10.1016/j.lindif.2013.03.011

*Wijekumar, K. K., Meyer, B. J. F., & Lei, P. (2012). Large-scale randomized control trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. *Educational Technology Research and Development, 60,* 986–1013. http://dx.doi.org/10.1007/s11423-012-9263-4

*Wijekumar, K., Meyer, B. J. F., Lei, P., Lin, Y., Johnson, L. A., Spielvogel, J. A., . . . Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for 5th-grade readers. *Journal of Research on Educational Effectiveness, 7,* 331–357. http://dx.doi.org/10.1080/19345747.2013.853333

*Wilkins, S. A. (2007). *Teaching expository text strategies to improve reading comprehension in low readers* (Unpublished dissertation). University of California, Riverside, CA.

Williams, J. P. (2005). Instruction in reading comprehension for primary-grade students a focus on text structure. *The Journal of Special Education, 39,* 6–18. http://dx.doi.org/10.1177/00224669050390010201

*Williams, J. P., Hall, K. M., & Lauer, K. D. (2004). Teaching expository text structure to young at-risk learners: Building the basics of comprehension instruction. *Exceptionality: A Special. Education Journal, 12,* 129–144.

*Williams, J. P., Hall, K. M., Lauer, K. D., Stafford, K. B., DeSisto, L. A., & deCani, J. S. (2005). Expository text comprehension in the primary

grade classroom. *Journal of Educational Psychology, 97,* 538–550. http://dx.doi.org/10.1037/0022-0663.97.4.538

*Williams, J. P., Nubla-Kung, A. M., Pollini, S., Stafford, K. B., Garcia, A., & Snyder, A. E. (2007). Teaching cause-effect text structure through social studies content to at-risk second graders. *Journal of Learning Disabilities, 40,* 111–120. http://dx.doi.org/10.1177/00222194070400020201

Williams, J. P., & Pao, L. S. (2011). Teaching narrative and expository text structure to improve comprehension. In R. E. O'Connor & P. F. Vadasy (Eds.), *Handbook of reading interventions* (pp. 254–278). New York, NY: Guilford Press.

*Williams, J. P., Pollini, S., Nubla-Kung, A. M., Snyder, A. E., Garcia, A., Ordynans, J. G., & Atkins, J. G. (2014). An intervention to improve comprehension of cause/effect through expository text structure instruction. *Journal of Educational Psychology, 106,* 1–17. http://dx.doi.org/10.1037/a0033215

*Williams, J. P., Stafford, K. B., Lauer, K. D., Hall, K. M., & Pollini, S. (2009). Embedding reading comprehension training in content-area instruction. *Journal of Educational Psychology, 101,* 1–20. http://dx.doi.org/10.1037/a0013152

Williams, J. P., Taylor, M. B., & deCani, J. S. (1984). Constructing macrostructure for expository text. *Journal of Educational Psychology, 76,* 1065–1075. http://dx.doi.org/10.1037/0022-0663.76.6.1065

Wilson, S. J., Tanner-Smith, E. E., Lipsey, M. W., Steinka-Fry, K., & Morrison, J. (2011). Dropout prevention and intervention programs: Effects on school completion and dropout among school-aged children and youth. *Campbell Systematic Reviews, 8,* 1–62.

Yochum, N. (1991). Childrens' learning from informational text - the relationship between prior knowledge and text structure. *Journal of Reading Behavior, 23,* 87–108.

Appendix A

Effect Size Equations and Considerations

Effect Sizes Estimation for True Experiments at the Student Level

For experimental studies, the standardized mean difference effect size (*d*) was used to represent intervention effects on the reading outcome measures identified for each study. The equations for the effect size and pooled standard deviation are represented as follows:

$$\mathrm{ES_{sm}} = \frac{\bar{X}_T - \bar{X}_C}{s_p} \quad \mathrm{s_p} = \sqrt{\frac{(s_1^2)(n_1 - 1) + (s_2^2)(n_2 - 1)}{n_1 + n_2 - 2}}$$

where ES_{sm} is the standardized mean difference effect size, $\bar{X}_T$ is the mean score of the treatment group on the posttest, $\bar{X}_C$ is the mean score of the control group on the posttest, and s_p represents the pooled standard deviation of the two groups. Positive scores favor the treatment group and negative scores favor the control group. In the equation deriving the pooled standard deviation, s_1 and n_1 represent the standard deviation and number of observed participants for the treatment groups, respectively, while s_2 and n_2 represent the same variables for the control group.

Effect Sizes for True Experiments at the Cluster Level

When clusters are the unit of random assignment and analysis, the effect size is not comparable with an effect size with assignment and analysis at the participant-level because it is calculated based on the between group variance instead of the total variance. To adjust for this, we calculated the effect size based on the between group variance and then multiplied the effect size by the square-root of the intraclass correlation:

$$d_T = d_B\sqrt{\rho}$$

where d_T is the standardized mean difference effect size based on the total variance, d_B is the effect size based on the total variance, and ρ is the intraclass correlation (Hedges, 2009). We estimated ρ for this equation as .20 in all instances (see the rationale in the subsequent section).

Effect Size Estimation for Quasiexperiments: Adjusting for Pretest Differences and Clustering

Some studies included in this meta-analysis used nested designs, in which classrooms were randomly assigned to treatment or control groups, but the analysis was conducted at the student level. When this occurred, the studies were classified as quasiexperiments and subject to an additional inclusion criterion (i.e., inclusion of a pretest measuring group differences prior to the intervention). Because an important function of randomization is to ensure a lack of assignment bias, failure to randomly assign participants increases the likelihood of inequalities between the treatment and control groups. Consequently, they were adjusted for pretest differences between the groups.

The authors computed the effect sizes (*d*) for these studies as the difference between the treatment and control condition (i.e., $\bar{Y} - \bar{Y}_{ctrl}$) after adjusting for pretest reading differences by subtracting the mean difference at pretest from posttest, or estimating the posttest mean-difference statistic from covariate-adjusted posttest means. This difference was then divided by the pooled standard deviation for the posttest. Although this has been a conventional approach to estimating effect sizes for quasiexperiments in previous meta-analyses (e.g., Bangert-Drowns, Hurley, & Wilkinson, 2004; Graham & Perin, 2007), more recent statistical approaches indicate using conventional methods to compute and average effect sizes across these studies may be inadequate due to incorrect variance estimation. In particular, Hedges (2009) provides

(*Appendices continue*)

various statistical models for estimating variance structures and calculating effect sizes in nested designs, illustrating the likelihood of underestimation of standard errors when using conventional statistics. To contend with this problem, Hedges recommends that meta-analysts choose a model to estimate or adjust these parameters in such a way that they are consistent and analogous to effect sizes of other studies to which the study will be compared. To this end, we chose to use the effect size for quasiexperiments as "ES = δ_T" (Hedges, 2009, pp. 340–343), as well as the corresponding standard error, so that each component was estimated based on a total variance that included both student and classroom level variance components.

However, in most cases the quasiexperimental studies in this review did not report appropriate data to calculate classroom level variance. Therefore, it was necessary to estimate δ_T by adjusting the conventional effect sizes using an intraclass correlation. Many of the studies included had equal sample sizes across clusters, and we assumed equal cluster sizes when the authors did not specify. Therefore, we chose to adjust effect sizes using the intraclass correlation estimator "ES = d_T" (Hedges, 2009):

$$d_T = \left(\frac{Y^T_{..} - Y^C_{..}}{S_T}\right)\sqrt{1 - \frac{2(n-1)\rho}{N-2}}$$

where $Y^T_{..}$ is the grand mean of the treatment group, $Y^C_{..}$ is the grand mean for the control group, S_T is the total pooled within-treatment variance, n is the number of students within cluster, N is the number of students total, and ρ is the intraclass correlation. In addition, we continued to adjust for possible pretest differences between the treatment and control groups by subtracting the mean difference at pretest from the mean difference at posttest.

The variance of the effect sizes also had to be adjusted to include the variance associated with clustering. The equation for calculating the variance of d_T is normally distributed, and was calculated using the following equation provided by Hedges (2009):

$$v_T = \left(\frac{N^T + N^C}{N^T N^C}\right)(1 + (n-1)\rho) + d_T^2\left(\frac{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}{2(N-2)[(N-2) - (N-1)\rho]}\right)$$

where N^T is the total number of students in the treatment group, and N^C is the total number of students in the control group (other symbols defined in the previous paragraph).

To use these formulas to adjust for clustering, it was necessary to impute the intraclass correlations (ICCs), or ρ, because they were also not reported in any of the source studies. To be conservative, we imputed an ICC value of .20 in all cases, regardless of grade level(s) of the study participants. This is consistent with the convention of What Works Clearinghouse, which has adopted an ICC value of .20 for achievement outcomes (What Works Clearinghouse, 2014).

Hedge's *g* (Small Sample Correction)

Finally, the standardized mean difference effect size (d) is upwardly biased in small samples (Hedges, 1981, as cited in Lipsey & Wilson, 2001). Therefore, a small sample correction was applied to the effect size to provide an unbiased effect using the following formula:

$$g = \left[1 - \frac{3}{4N-9}\right]d$$

where g is the small sample correction of the standardized mean difference effect size (d) and N is the total sample size. As the total sample size increases for a particular study, the correction to the effect size becomes negligible. Therefore, the correction was applied to all effect sizes.

Combining Data Across Multiple Treatments or Multiple Subgroups Within a Treatment

In some of the studies included in this review, researchers compared multiple grade levels or multiple student types (e.g., below average, average, students with learning disabilities) within conditions. In these instances, it was sometimes necessary to aggregate the performance of two or more groups as a prelude to calculating the effect size for some comparisons. To aggregate data within each condition, the procedure recommended by Nouri and Greenberg (1995) was applied (Cortina & Nouri, 2000). This procedure estimates an aggregate group or grand-mean. We first calculated the aggregate treatment or control mean as an n-weighted average of subgroup means:

$$\bar{Y}_{..} = \frac{1}{n_{..}}\left[\sum_{j=1}^{k}(n_{.j})(\bar{Y}_{.j})\right]$$

where $\bar{Y}_{..}$ is the grand mean, $n_{..}$ is the total number of students within the condition, k is the number of groups within the condition, $n_{.j}$is the number of students in the jth group within the condition, and $\bar{Y}_{.j}$ represents the mean for the jth group within the condition.

Next, the aggregate variance was calculated by adding the n-weighted sum of squared deviations of group means from the grand mean to the sum of squared deviations within each subgroup:s

$$s^2_{..} = \frac{1}{n_{..} - 1}\left[\sum_{j=1}^{k} n_{.j}(\bar{Y}_{..} - \bar{Y}_{.j})^2 + \sum_{j=1}^{k}(n_{.j} - 1)s^2_{.j}\right]$$

where $s^2_{..}$ is the total variance for the condition, and $s^2_{.j}$ is the variance for the jth group within the condition, with all other variables defined in the previous paragraph.

Aggregated treatment or control means and standard deviations were used when computing an overall independent effect size (d) for each study in the analysis of the three main research questions.

(*Appendices continue*)

Aggregating Across Multiple Measures

Across studies, there was no single reading measure used by a majority of investigators. For example, researcher-devised measures of reading comprehension included answering questions about text (multiple choice and short answers), retelling what was read (orally or in writing), summarizing text read in one sentence, and identifying words systematically omitted from text (cloze procedure). As a result, there was no single assessment that could be used as the sole measure of reading comprehension. Moreover, many researchers administered multiple tests of reading comprehension. Consequently, effect sizes for multiple measures within a study were aggregated using a simple average. Aggregation of effects across different measures for the same construct is preferable when intercorrelations among the measures are unknown, as standard error estimation is complicated when this information is missing (Gleser & Olkin, 1994).

Appendix B

Sensitivity Analysis to Examine the Impact of the Assumed Value of Rho on the RVE Meta-Analysis for RQ1a

rho	*k*	ES	*SE*	95% CI	*df*[1]	*p* value	τ^2
.0	40	.5742	.0910	[.39, .76]	31.89	<.001	.1507
.2	40	.5742	.0910	[.39, .76]	31.90	<.001	.1508
.4	40	.5742	.0910	[.39, .76]	31.90	<.001	.1509
.6	40	.5742	.0910	[.39, .76]	31.90	<.001	.1510
.8	40	.5743	.0910	[.39, .76]	31.91	<.001	.1511
.99	40	.5743	.0910	[.39, .76]	31.91	<.001	.1512

Note. RVE = robust variance estimate; RQ1a = Research Question 1a; *k* = number of studies; ES = effect size; CI = confidence interval.
[1] Degrees of freedom reflect small sample adjustments suggested by Tipton (2015).

Appendix C

Publication Bias Analyses

Because we had only a small number of studies in our primary analysis (RQ1a, $k = 40$), we conducted four analyses to test for potential publication bias. First, we analyzed using a funnel plot (see Figure C1).

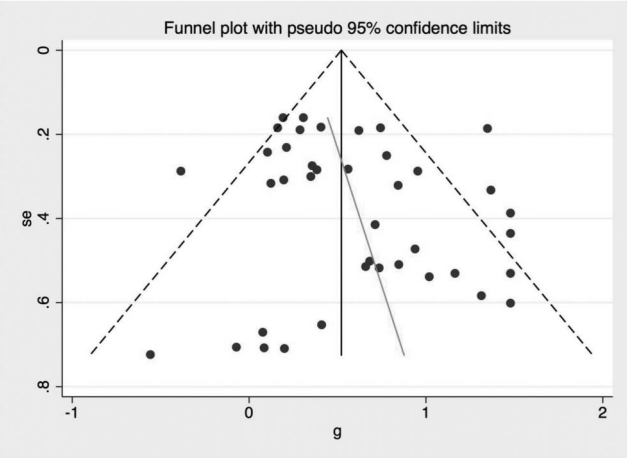

Figure C1. Funnel plot with 95% confidence interval for RQ1a.

The funnel plot seems to have a slight lack of symmetry, indicating that some studies with null or negative findings have not been found and included in our analysis. The funnel plot also seems to indicate studies with the smaller sample sizes, and larger standard errors, were more likely to have smaller or negative effects. Second, we conducted Egger's test for small study effects, providing a regression-based approach that regresses the intervention effect estimates on their standard errors. However, the results of the test indicated there were no small-study effects ($p = .174$). Third, we conducted a failsafe *N* analysis, which indicated 154 additional studies would need to be found with a null result in order to render the average weighted ES nonsignificant. Finally, in an exploratory trim and fill analysis (Duval & Tweedie, 2000), four effect sizes were trimmed and filled. Using a random effects model, the estimated ES was reduced from 0.57, 95% CI [0.42, 0.72] to 0.51, 95% CI [0.35, 0.66]. The estimate of the ES provided in the trimmed and filled analysis is slightly more conservative, but confidence intervals did not cross zero in either case, and overlapped to a large degree.

Received September 4, 2014
Revision received August 14, 2015
Accepted August 16, 2015 ■