Sources of variance in reading comprehension research:

The role of measures and interventions

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Calvary Diggs

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. Theodore J. Christ, Adviser

May 2020

Abstract

The purpose of this study was to examine if differences in reading comprehension measures' response formats were associated with differential outcomes for reading comprehension interventions. Specifically, this study used meta-analysis to evaluate the overall treatment effect of reading comprehension interventions, the association between a measure's response format and measured intervention outcomes, and whether specific intervention effects varied based on the measure's response format. A systematic review of the literature identified 66 published and unpublished research reports and studies conducted since 2000. All studies administered a reading comprehension intervention for students in the primary grades and measured the effects using a reading comprehension measure. Meta-analytic findings suggested an overall positive effect of reading comprehension interventions for both intervention to control group comparisons at posttest (Hedge's $g = 0.20$) and pretest to posttest comparisons in the intervention group (Hedge's $g = 0.71$). The response format of a reading comprehension measure, specifically retell/summary formats, was significantly associated with intervention outcomes, even after controlling for purposively selected variables. Findings also indicated that improving background knowledge and multicomponent interventions were significantly associated with performance on measures of reading comprehension with retell/summary response formats. The results of this study provide additional evidence that measures using the retell/summary response formats value reading comprehension differently, specifically in the context of interventions. Findings may also be used to

caution against the interchangeable use of retell/summary formats with other measures of reading comprehension.

Keywords: reading comprehension, meta-analysis, measure, intervention

Table of Contents

List of Tables iv

List of Figures v

Chapter 1: Introduction 1

Chapter 2: Literature Review 9

Chapter 3: Method 39

Chapter 4: Results 57

Chapter 5: Discussion 98

Bibliography 120

Appendix A 149

Appendix B 152

Appendix C 155

Appendix D 157

List of Tables

Table Page

1. Study-level Descriptive Characteristics for Included Studies 58
2. Study-level Assessment Characteristics 60
3. Study-level Intervention Characteristics.....62
4. Sensitivity Analyses for Posttest and Growth Meta-analyses.....70
5. Study Design Meta-regressions 73
6. Intervention Meta-regressions 74
7. Assessment Meta-regressions 76
8. Reader Characteristic Meta-regressions 77
9. Study Quality Meta-regressions.....78
10. Full Model Meta-regression Using all Relevant Block Variables 86
11. Retell/Summary-only Meta-regression with Significant Intervention Variables ..95

List of Figures

Figure Page

1. Flow diagram of study search and inclusion procedures 43
2. Funnel plot of all posttest effect size estimates compared to their standard errors.. 64
3. Funnel plot of all within-group growth effect size estimates compared to their standard errors. 65

Chapter 1: Introduction

It is important that individuals acquire the necessary skills and abilities to read with comprehension. The National Reading Panel ([NRP] 2000) conducted an extensive review of scholarly literature in reading. They explicitly stated that although it is important to develop basic skills in reading, the purpose of reading is comprehension. Indeed, a major focus in early schooling is teaching students the skills to read, and a later focus is using those reading skills to learn new information (Chall, 1993). One way that schools understand student reading development is through multi-tiered systems of support (MTSS) and response to intervention (RTI). Under those frameworks, schools utilize systemwide screening to identify reading problems. Ideally, the tools used to screen and monitor reading problems have evidence to support their interpretation and use for those purposes (Christ & Nelson, 2014). Schools also provide tiered supports and services to further monitor and intervene on specific problems in reading. This framework would suggest that about 20% of students in a system require additional supports above and beyond core instruction (Schulte, 2016). The supports should also be practices that have prior evidence-base supporting their effectiveness. For example, What Works Clearinghouse, which is a clearinghouse for education research, has 40 intervention programs to-date related to reading comprehension. A total of 30 of those interventions have convincing or promising evidence of their effectiveness, while each intervention varies in the extent to which it directly teaches reading comprehension skills (What Works Clearinghouse [WWC], 2019). Altogether, there are variety of measures

and interventions that could be used to support the reading comprehension needs of students; however, challenges do exist.

Statement of the Problem

Many students in the United States experience some difficulty in reading. On the 2019 National Assessment of Education Progress (NAEP), only 35% of fourth grade students and 34% of eighth grade students scored at or above proficient in reading. Moreover, these scores were significantly lower than the previous measurement period (U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, NAEP, 2019). School problems in reading are associated with general disadvantages in the school context (Hudson, Torgesen, Lane, & Turner, 2012) and with later school dropout when performance is below third grade standards (Alexander, Entwisle, & Kabbini, 2001). Thus, acquiring the skills to read and accurately applying those skills is critical.

Fortunately, there is much research in reading, including how to identify and support students experiencing reading problems (e.g., Ball & Christ, 2012; Burns et al., 2016; Fuchs & Fuchs, 2006; Jenkins, Hudson, & Johnson, 2007; Joseph, 2015; Neddenriep, 2014). Teaching readers comprehension strategies are among the most common reading comprehension interventions; such strategies include, question-generation, activating prior knowledge, and identifying the main idea (Joseph, 2015; Neddenriep, 2014; Sencibaugh, 2007; Shanahan et al., 2010). Other interventions in reading comprehension, which are also defined as strategies, include summarizing, instruction in inferencing, and self-monitoring (Duke & Martin, 2015; Joseph, 2015;

Neddenriep, 2014). In school-based applications, the use of measures and interventions designed for detecting and supporting students with reading problems serve, in part, as the foundation to current theoretical orientations to universal prevention, identification, and tiered remediation efforts in reading (Fuchs & Fuchs, 2006). However, critical syntheses and evaluations of the research are still needed, for there are still issues regarding the selection, interpretation, and use of data to inform schoolwide prevention and intervention efforts (Dailor & Jacob, 2011).

For example, there is a paucity of critical syntheses of the research and evidence regarding the reading comprehension measures used to identify and monitor reading comprehension interventions. Outside of measures used for summative and screening purposes, reading comprehension measures may be used to monitor the progress of interventions (Salvia, Ysseldyke, & Bolt, 2007). The National Center on Intensive Intervention (2012) reviewed a variety of academic progress monitoring measures, including two in reading comprehension. These progress monitoring measures, such as mCLASS (e.g., Snow, Morris, & Perney, 2018) and easyCBM (e.g., Lai, Irvin, Alonzo, Park, & Tindal, 2012), can be used to measure student growth in reading comprehension over time as well as the effect of intervention or instruction. Although measures of reading comprehension are used to monitor student response to interventions, there are clear distinctions between measures. Two such distinctions are their response formats and the validity evidence supporting their interpretation and use (see Chapter 2). It is currently unclear if the varying response formats of the measures relates to variation in the efficacy of reading comprehension interventions. In other words, research is needed

to investigate if reading comprehension measures could and should be used interchangeably to monitor intervention effects.

To summarize the current example and reason for this dissertation: there is a gap in the current research regarding the interchangeability of reading comprehension measures as they relate to measuring intervention effects. Findings regarding whether reading comprehension measures are interchangeable has the potential to influence both research and practice. This study aims to help address an existing gap in the literature regarding the use of reading comprehension measures to monitor intervention effects. The study synthesizes current and available research on reading comprehension interventions and measures. Findings inform current knowledge and considerations regarding the selection and use of reading comprehension measures in both research and practice.

This study is important because the results describe and analyze the influence of measurement on intervention effects reported in research literature. Specifically, this study explores the role of the measure in moderating intervention effectiveness. Current evidence suggests that reading comprehension measures assign values to comprehension performance differently (Collins, Lindstrom, & Compton, 2018; Garcia & Cain, 2014; Kendeou, Papadopoulos, & Spanoudis, 2012; O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014), which may influence the measured effect. Although evidence of these differences exists, published meta-analyses of reading comprehension often aggregate effects measured with alternate dependent variables as though they are interchangeable (e.g., Edmonds et al., 2009; Peng et al., 2018; Scammacca, Roberts, and Stuebing, 2014;

Swanson et al., 2017). Given that measures are used interchangeably, the influence of using different dependent measures to value the effects of interventions should be examined to inform future syntheses and interpretations of meta-analytic findings in reading comprehension. As such, the impetus for the current research is to seek further knowledge regarding whether the methodological choices in the measure used influences the observed effects of interventions in reading comprehension.

Relevant Definitions

The purpose of the current section is to introduce relevant definitions associated with the present dissertation. To begin, *Reading comprehension* is the culmination of behavioral processes, which include tracking and decoding, that have become automatized and linked to language and cognition to interact with the explicit and implicit content of written language to develop a cohesive mental representation (Magliano, Millis, Ozuru, & McNamara, 2007; Pawlik & Rosenzweig, 2000; Sabatini, Albro, & O'Reilly, 2012). Next, *Intervention* is defined by WWC (2019) as, "An educational program, product, practice, or policy aimed at improving student outcomes."

Reading comprehension interventions were defined using Suggate's (2016) criteria for reading interventions, which first denotes other types of reading interventions before defining a reading comprehension intervention:

> Phonemic awareness (and phonological awareness) interventions focused on manipulation of sounds in the absence of text and phonics included letter–sound or sound–spelling relations. Fluency interventions focused on skill at reading connected text, to the exclusion of practice at reading sentences or single words

> (e.g., peer tutoring, repeated reading). Comprehension interventions were those that focused on strategies to decipher text and derive meaning without a phonics focus, such as summarizing, prior knowledge, and inferential thinking (p. 82).

Thus, reading comprehension interventions can be described as the process of teaching students strategies, skills, or practices that support reading comprehension. In intervention, students learn to use strategies intentionally or automatically and frequently or infrequently to achieve a particular goal while reading, which typically relates to their understanding and application of the text (Suggate, 2016; van den Broek, Beker, & Oudega, 2015). Suggate's (2016) conceptualization was used because it provided broad, yet clear definitions of multiple intervention types (i.e., strategy instruction, text structure instruction, improving background knowledge, self-monitoring, inference instruction, graphic organizers, multicomponent; Joseph, 2015; Little & Akin-Little, 2014) which would allow identification of interventions that utilized multiple components in addition to reading comprehension. In addition to the content of interventions, their formats are quite broad and may be viewed as intentional efforts to prevent or remediate problems in school settings (Kratochwill, Clements & Kalymon, 2007). Interventions supplement the universal instruction and supports that all students receive as part of being in a school (Burns, Riley-Tillman, & Rathvon, 2017).

Study Purpose and Significance

The purpose of this study was to examine the link between measurement, intervention, and the measured effects of interventions for students in the primary grades. The specific focus was reading comprehension and not prerequisite and concurrent skills,

such as oral language, phonological awareness, phonics, vocabulary, or fluency. Moreover, this study focused on reading comprehension interventions implemented in English for elementary-aged students in the United States. This was done so that the findings were targeted and specific to a particular context, rather than broad and amorphous.

Meta-analytic methods were used to evaluate (a) group differences at posttest for students who received reading comprehension intervention, (b) pretreatment-to-posttreatment change, and (c) the extent to which intervention outcomes differed based on the dependent variable used to measure treatment effects. In other words, this study examined intervention effects and the influence of measurement on those effects. It was expected that effect sizes would vary based on the reading comprehension measure used to monitor intervention effects.

Research Questions

Three research questions (RQ) are the main focus of this dissertation and are explored using meta-analysis:

> RQ 1: In standardized mean difference, as measured through Hedge's *g*, what is the observed aggregate treatment effect for all reading comprehension interventions delivered to students in elementary school at posttest compared to control?

RQ 2: In standardized mean difference, as measured through Hedge's *g*, what is the observed aggregate treatment effect for all reading comprehension interventions delivered to students in elementary school from pretest to posttest?

RQ 3: Do observed treatment effects differ based on the specific interventions and measures?

Chapter 2: Review of Literature

The purpose of the second chapter is to provide a review of relevant literature. The chapter is organized into four broad sections. The first section describes reading comprehension and how it is assessed. It also introduces the argument-based approach to validity and its relationship to reading comprehension assessment. The section that follows briefly describes reading comprehension interventions. Next, the paper reviews relevant meta-analyses in reading comprehension to understand the extent to which the relationship between measurement and interventions has been investigated. The final section is a critical review and synthesis of reading comprehension measures to further highlight why the specific measure used to assess intervention outcomes may be an important source of variance.

Interpretation and Use of Comprehension Measures

Reading assessment is used to understand student and system-level performance in the domain (e.g., Ball & Christ, 2012; Jenkins et al., 2007). Data-based decision-making with assessment data can also help improve school-based practices in reading (e.g., Burns et al., 2016). However, schools must collect the right data using the proper measure to provide meaningful information, which is no simple task (Dailor & Jacob, 2011).

Reading is multifaceted, lending itself to the development of measures in various areas. For example, alphabetic knowledge, decoding, vocabulary, fluency, comprehension, prior knowledge, and metacognitive strategies are all broad factors

related to reading that can be assessed in broad or narrow ways (e.g., Amendum, Conradi, & Pendleton, 2016). Of note, early literacy (e.g., phonological awareness) and the application of basic skills in reading (e.g., fluent decoding) represent a comprehensive literature of meaningful components of reading (e.g., Hammill, 2004). Helping students acquire early literacy and fluency is a necessary step in reading for understanding (NRP, 2000). However, fluency, for example, should not be the only area of reading measured in schools, yet these measures are widely used in schools adopting a data-based decision-making model (Carson, 2017; Fuchs, & Fuchs, 2006). Research can still help improve measures in early and basic reading skills to benefit educational practices; however, integrating reading comprehension measures into school data-based decision-making systems is still needed (Fuchs & Vaughn, 2012; Sabatini et al., 2012). Assessing reading comprehension is crucial, for comprehension is the purpose of reading (NRP, 2000).

Assessing reading comprehension is complex since different measures assess different aspects of comprehension (Kendeou et al., 2012). For example, reading comprehension may be described and measured as awareness of the chain of events of a story, simple recall, identifying the main ideas of the text, answering questions, and applying the information to complete a task (e.g., van den Broek, 1988; van den Broek et al., 2005).

A key premise of the current chapter is that reading comprehension measures are important components of school-based formative, diagnostic, and evaluative assessment practices; however, useful reading comprehension measures for school-based contexts have not been properly identified in a way that facilitates effective selection and use. The

current chapter seeks to synthesize evidence in the peer-reviewed literature supporting the interpretation and use of existing reading comprehension assessments in school data-based decision-making systems. In order to provide a meaningful synthesis, it is necessary to examine relevant validity evidence.

Unified validity arguments in assessment. Measures are used to gather information that may or may not be useful based on the purpose. For example, a measure of oral language comprehension may not provide useful information about how well a student understands written passages, but a set of open-ended questions might. Some broad sources of validity evidence are psychometrics, content, educational usability, and reduction of construct irrelevant variance (CIV). They can be used to support the use of measures for common purposes in schools such as screening or progress monitoring. The current section provides some parameters regarding different forms of validity evidence gathered from the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014).

Psychometric evidence is the set of statistical and descriptive properties of a test used to describe a construct. Reliability of scores and validity of a particular interpretation of those scores are core features of psychometric evidence. Reliability indices describe a score's consistency for a group or a single individual. It also expresses the degree to which a score may be affected by measurement error from one time point to the next. A valid (i.e., legitimate) score interpretation depends on reliable (i.e., consistent) scores. A typical rule is that a reliability of at least .80 is needed to make decisions about

individuals (Thorndike & Thorndike-Christ, 2011). Typical indices of reliability are test-retest reliability, split half reliability, alternate form reliability, and standard error of measurement. These are all expressed in terms of numbers and coefficients.

Unlike reliability, validity can be expressed in terms of quantitative metrics and qualitative appraisal. Validity evidence may be defined holistically or narrowly, given the context and uses of the score. The relation of one measure to other measures of importance (e.g., criterions) are sources of validity evidence.

Another source of validity evidence is the content of the measure. Content-related validity evidence is assessed through empirical (e.g., factor analysis, which also provides evidence of the internal structure of a test) and qualitative means (e.g., test blueprint, expert judgment). It is the extent to which the items of the measure are related to the construct it is supposed to represent. For example, a measure might assign a value based on knowledge of setting by asking questions about the location and time period of the story.

Educational usability is defined here using features of Deno's (1985) criteria for curriculum-based measurements (CBM). In essence, educational usability is the extent to which a test is easily adopted and used within the educational environment. Factors that may affect this are (a) the time to administer the test per pupil, (b) the training required to administer and score, (c) the usefulness of those data for informing instructional decision making, (d) the cost of the measure, and (e) capacity for diagnostic and growth assessment.

Finally, assessing reading comprehension is accompanied with domain-specific challenges. Though not comprehensive, some of the features to consider are the role of prior knowledge, attitudes and motivation, and differential item functioning and scorer behavior. These can be generally labeled as CIV.

Conclusion: Unified validity arguments in reading comprehension. Altogether, these forms of evidence can be used to evaluate the interpretation and use argument (IUA) for various reading comprehension assessment scores. Notably, the validity evidence gathered varies based on the nature of the IUA, whereby specific evidence is used to address the elements of the IUA. This can be conceptualized under Kane's (2013) argument-based approach to validity. The approach is that the target of validation is the IUA, rather than the measure or its scores, which is done through an evaluative process. This requires at least two steps: (1) a clear statement of the purported IUA (i.e., a proposal) and (2) an evaluation of the evidence to support such claims (i.e., examine plausibility). Validation occurs when the IUA is clear and the validity argument is plausible based on evidence; however, the available evidence must relate to the IUA. Of note, Cook, Brydges, Ginsburg, and Hatala (2015) and Kane (2013) provide an in-depth review and application of the approach. In relation to the present research, a score attained from a reading comprehension measure must have evidence to support its use in K-12 settings. The evidence needed (e.g., usability, psychometric evidence, content-related validity evidence) varies based on the claims being made (e.g., the assessment can feasibly be administered in a school setting for screening). As such, the evidence may vary, but evidence is always needed to support an IUA. For example, one interpretative

argument is that reading comprehension measures may be used to monitor the effects of reading comprehension interventions. The section that follows describe common reading comprehension interventions.

Reading Comprehension Interventions

The purpose of this section is to provide a general overview of reading comprehension interventions used in elementary school settings. As described in the first chapter, an intervention is a program, practice, policy, or product used with the intent to improve student outcomes (WWC, 2019). Likewise, reading comprehension interventions are interventions that incorporate reading comprehension practices in their methods and aim to improve student text comprehension (Suggate, 2016). Reading comprehension was not formally taught before 1980, but research began to investigate the formal instruction in reading comprehension in the 80s outside of genre-specific discourse (NRP, 2000).

Reading comprehension interventions are often identified broadly under the domain of reading comprehension strategy instruction (Neddenriep, 2014). Reading comprehension strategies are broad and diverse in nature, but they may be succinctly described as procedures used to support readers in understanding text before, during, and after reading (Neddenriep, 2014; NRP, 2000). Comprehension strategies are modeled by the instructor until the student can independently use the strategy without the support of the teacher (NRP, 2000). In a review of reading comprehension interventions, Joseph (2015) summarized a range of reading comprehension strategies: Ask, read with alertness, tell (ART); Embedded story structure (ESS); FIST (question generation and

reflection); What I know, what I want to know, and what I learned (KWL); Paragraph shrinking; Question-and-answer relationship (QAR); Read, ask, paraphrase, question (RAP-Q); Reciprocal teaching; Response cards; Survey question, read, recite, review (SQ3R); and Think before reading, think while reading, and think after reading (TWA). Overall, these approaches contain activating prior knowledge, self-questioning, identifying main idea, paraphrasing and retelling, as well as summarizing. In this way, reading comprehension interventions may include one or multiple strategies.

The NRP (2000) reiterates that reading comprehension is a complex process, where readers may approach written text with various purposes (Sabatini et al., 2012). In order to effectively navigate these purposes, readers must approach the text using their own background knowledge to derive meaning from the text in order to fulfill their reason for reading. Joseph (2015) described the process of activating prior knowledge as the reader engaging in how their current knowledge and experiences relate to the passage in front of them.

Given their review of the research, the NRP (2000) found that *comprehension monitoring* (the reader learns to be aware of their understanding of the text as well as what to do when problems negatively affect comprehension), *cooperative learning* (readers develop strategies with peers while reading), *graphic organizers* (visual and written representations of content from the text), *story structure instruction* (learning to ask questions and identify information relevant to the plot or text genre), *question answering* (providing answers to questions presented by the instructor and provided feedback), *question generation* (reader asks themselves questions to support

comprehension, typically beginning with who, what, when, why, where, and how), *summarization* (distilling the ideas in a text into a coherent whole or identifying the main idea or key details), and the teaching of *multiple strategies* (the flexible use of multiple strategies and procedures, while consulting the teacher) were effective procedures supported by evidence from 203 studies on text comprehension.

In addition, Joseph (2015) described paraphrasing and retelling, in addition to summarization. The distinction that Joseph (2015) made was that summaries synthesize information, paraphrasing restates information in the reader's own words, and retelling is the process of recalling key details from the passage. In this way, paraphrasing is the least complex, while summarizations are the most complex, given that they require the skills used in paraphrasing, retelling, and making connections between different ideas which may require the use of inferencing and activating prior knowledge.

Inferencing was referenced in the previous paragraph. In the context of reading, readers make inferences when they use information other than what was explicitly stated in the text to understand the passage (van den Broek et al., 2015). Inference-making is related to reading comprehension (Kendeou, McMaster, & Christ, 2016). Inference instruction is a way to teach students to understand and identify the implicit ideas and concepts within the text (Elleman, 2017). Inference instruction incorporates some of the reading comprehension strategies discussed in the previous paragraphs (e.g., question generation, activating prior knowledge; Kendeou et al., 2016). Inference instruction supports readers in incorporating their own knowledge and experiences as well as information from within the text and other texts to understand the passage (Elleman,

2017; Kendeou et al., 2016). Similar to strategies, inferences are made at the readers' discretion and initially with systematic support from the instructor (Kendeou et al, 2016; Sabatini et al., 2012).

Altogether, there are a variety of ways and methods to improve reading comprehension. They may be broadly defined as strategies; however, they may also be more narrowly defined. The next section describes current knowledge from meta-analyses regarding the relationship between intervention and measurement in reading comprehension.

Sources of Variance in Meta-analyses on Reading Comprehension

Meta-analysis is a quantitative approach to research synthesis that aggregates statistical data (e.g., correlations, means, standard deviations) to summarize findings in a body of research (Card, 2012). It is considered the highest level of evidence because it quantitatively synthesizes a research domain (Cooper, Hedges, & Valentine, 2009). A number of meta-analyses were identified that have been conducted in reading comprehension. This section broadly summarizes findings from those meta-analytic reviews regarding sources of variance in reading comprehension. This section also explores factors related to reading comprehension that have been explored through meta-analysis, as they relate to reading comprehension assessment.

A total of 39 meta-analyses on reading comprehension were reviewed for sources of variance in reading comprehension (Araujo, Reis, Petersson, & Faisca, 2015; Berkeley, Kurz, Boykin, & Evmenova, 2015; Berkeley, Scruggs, & Mastropieri, 2010; Brown, Oram-Cardy, & Johnson, 2013; Collins et al., 2018; Edmonds et al., 2009; Ehri,

Nunes, Stah et al., 2001; Ehri, Nunes, Willows et al., 2001; Elleman, 2017; Florit & Cain, 2011; Follmer, 2018; Garcia & Cain, 2014; Graham & Hebert, 2011; Guthrie, McRae, & Klauda, 2007; Haller, Child, & Walber, 1988; Hebert, Bohaty, Nelson, & Brown, 2016; Hebert, Gillespie, & Graham, 2013; Kaldenberg, Watt, Therrien, 2015; Kim & Quinn, 2013; Kovachy, Adams, Tamaresis, & Feldman, 2014; Lee & Shu-Fei, 2017; Li, 2014; Lietz, 2006; Melby-Lervag, Redick, & Hulme, 2016; Moran, Ferdig, Pearson, Wardrop, & Bomeyer, 2008; Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009; Neville & Searls, 1991; Peng et al., 2018; Readance & Moore, 1981; Scammacca, Roberts, Vaughn, & Stuebing, 2015; Sencibaugh, 2007; Shenderovich, Thurston, & Miller, 2016; Spencer & Wagner, 2017; Spencer & Wagner, 2018; Suggate, 2016; Swanson et al, 2017; Swanson, 1999; Tran, Sanchez, Arellano, & Swanson, 2011; Wood, Moxley, Tighe, & Wagner, 2018). Although sources of variance differed across meta-analyses, common broad areas were interventions (*k* studies = 17; e.g., sentence-combining, inference instruction, classwide discussion), assessments (*k* = 13; e.g., timed vs. untimed, standardized vs. experimental, response format), learner characteristics (*k* = 14; e.g., disability status, grade level, executive function), other reading skills (*k* = 4; rapid automatized naming, listening comprehension, nonsense word reading), instructional accommodations and modifications (*k* = 3; i.e., technology), and most studies explored characteristics of the studies included in the meta-analysis (e.g., study design, randomization, treatment-control equivalence at pretest).

Reading comprehension interventions were primarily discussed in terms of aggregate effect sizes. The degree of specificity varied by study. For example, some

studies described broadly all interventions as reading interventions, comprehension interventions, or fluency interventions (e.g., Scammaca et al., 2015; Suggate, 2016; Swanson et al., 2017). Other studies described types of comprehension strategies within comprehension interventions, for example, and provided effect sizes for each of the strategies (e.g., Hebert et al., 2016). Very few meta-analyses described intervention effects for discrete intervention packages (e.g., Guthrie et al., 2007). A similar pattern was true for measures.

Reading comprehension measures were discussed similar to interventions. It was most common for reading comprehension measures to be described in terms of standardized and experimental (e.g., Lee & Tsai, 2017; Murphy et al., 2009; Swanson et al., 2017) or simply in terms of the construct it was designed to measure (e.g., Peng et al., 2018 [Comprehension]). As such, most meta-analytic reviews of reading comprehension interventions have failed to investigate the connection between the measures and intervention outcomes beyond the acknowledgement that standardized, criterion-referenced measures of reading comprehension tend to result in lower effect sizes than experimentally-designed measures (Scammacca et al., 2015; Willingham, 2007). This is a broad distinction that does not account for differences in measures beyond standardization.

Meta-analyses were further examined for analyses that examined how individual assessments procedures moderated intervention effects. Only three studies performed analyses somewhat related to this topic. Hebert and colleagues (2013) hypothesized that the alignment between an intervention and its measure would result in greater effect sizes

for intervention in cases where the measure and intervention were aligned compared to interventions where the measure was not aligned. Specifically, they examined this in relation to whether the measures and interventions included a writing component. Overall, they found that there were higher effects. Hebert et al. (2013) examined the effects between alignment between assessment and intervention and found that alignment was a significant source of variation between effect sizes. In reading comprehension, it was the only meta-analysis to conduct such an investigation. Others performed variations that were less aligned to intervention. Instead they focused primarily on assessment.

Collins et al. (2018) and Garcia and Cain (2014) conducted meta-analyses that were assessment focused. They examined how differences in measures related to the reading comprehension performance of students based on theoretical relationships between decoding and comprehension (Gough & Tunmer, 1986; Garcia & Cain, 2014) as well as differences in performance of typically developing students and students with reading comprehension difficulties (Collins et al., 2018). Based on the variation in effect sizes, findings from both studies supported that measures of reading comprehension differentially relate to aspects of the domain. Unfortunately, a gap in the evidence still exists. For example, it is unclear if there are combinations between measure and intervention that result in differential outcomes, as there were in the context of writing-based assessments and interventions of reading comprehension (Hebert et al., 2013).

Collins et al. (2018) explored specific response formats of reading comprehension measures (i.e., cloze, multiple choice, SVT, open-ended questions, retell, and picture selection) and their relationship to students with and without reading difficulties. They

found large effect size differences between students with and without reading difficulties, where students with reading difficulties scored lower. This study showcases that assessments do, in fact, differentially value performance. However, this study only investigated differences between typically developing students and those with reading difficulties. It did not investigate interventions. Thus, the degree to which specific assessment formats for reading comprehension influence the magnitude of intervention outcomes is still unknown and requires additional investigation.

In conclusion, it appears that many of the reviews (k = 17) discussed student and intervention characteristics that influenced reading comprehension performance. However, only three meta-analyses (Collins et al., 2018; Garcia & Cain, 2014; Hebert et al., 2013) investigated how different types or specific measures influence reading comprehension scores. Furthermore, no studies investigated the link between specific measures and interventions. The next section reviews how reading comprehension measures vary in validity evidence and how that has potential to influence decisions about a student's reading comprehension in relation to the assessment-to-intervention process.

An Examination of Validity Evidence of Various Reading comprehension measures

Based on the above standards for assessment validation, reading comprehension measures were examined for validity evidence supporting their interpretation and use in k-12 settings. Specifically, over 50 studies in the peer reviewed research literature were consulted to understand extant evidence of (a) psychometric properties, (b) educational usability, (c) content-related validity evidence, and (d) reduction of CIV. In total, seven

assessment procedures were examined. The next sections describe each of the procedures and, as a case example, the current evidence supporting their IUA for screening and progress monitoring. Screening and progress monitoring are discussed, given that they are common purposes of assessment relevant to schools (National Center on Response to Intervention [NCRTI], 2012; Salvia et al, 2007).

Cloze unified validity evidence. A total of seven studies (six articles) provided validity evidence on the IUA of cloze scores for students in K-12 (Ashby-Davis, 1985; Berk, 1979; Fuchs & Fuchs, 1992; Fuchs, Fuchs, & Maxwell, 1988; Jones & Pikulski, 1979; Smith & Zinc, 1977). Of those studies, three explicitly described how cloze relates to reading comprehension: cloze measures a reader's understanding of the author's language. Thus, cloze measures reading comprehension and writing. Cloze functions as a passage with words that have been systematically deleted (e.g., every 5th word), and students restore the deletions (typically) by providing exact replacements. The number of words correctly replaced are tallied and scored based on various methods.

Cloze has promising evidence to support its IUA for screening reading comprehension problems. Criterion-related validity evidence suggests cloze has sufficient relationships with measures of reading (Fuchs et al., 1988; Jones & Pikulski, 1979; Smith & Zinc, 1977). In addition, the measure can be group administered in a relatively small amount of time (i.e., under ten minutes). Scoring may take longer and have weaker interscorer agreement depending on whether total exact replacements or a different method is used (Fuchs et al., 1988). In addition, no studies investigated classification accuracy, resulting in a lack of evidence of the consistency between low performance on

the cloze and concurrent low performance on a criterion measure, which would be valuable information regarding it's utility as a screener. In addition, no standard materials with norms and benchmarks appeared in the literature.

There was no evidence to suggest that cloze has been used across large (Mdn = 53.5; Range 30-70; total studies (k) = 4) or particularly diverse samples. Fuchs et al. (1988) was the only study to report demographic information for the sample, where 31% of the participants were students of color and all students received special education services. Cloze was used with student populations from late elementary school through high school (i.e., grades 4 to 11; k = 4).

Collectively, the existing evidence in the published peer-reviewed literature (k = 7) cautions the IUA of cloze in K-12 settings. Criterion-related validity evidence is promising. However, there was no evidence of test-retest and alternate form reliability, and internal consistency did not provide convincing evidence at or above the .80 standard (KR-21 = .76; Smith & Zinc, 1977). This would be a major concern for progress monitoring and screening, given that assessments must first be reliable before they can be used and interpreted validly (Kane, 2006). In addition, both content-related validity evidence and CIV raise concerns. Cloze requires skill in composition (Ashby-Davis, 1985), which could result in differential scores for students that are unrelated to reading comprehension (Fuchs & Fuchs, 1992). Furthermore, special education teachers preferred to use other measures because of the strict scoring procedures (Fuchs et al, 1988). Finally, the cloze has not been investigated with linguistically diverse samples. As such, there are improvements to make to the cloze regarding its standardization, acceptability

by students and teachers, reliability evidence, and understanding its relationship with reading comprehension from not only a psychometric lens but also a content perspective.

Informal reading inventories unified validity evidence. A total of ten articles (11 studies) provided validity evidence on the IUA of informal reading inventory (IRI) scores for students in K-12 (Clark, Kamhi, Nippold, & Boudrea, 2014; De Santi & Sullivan, 1984; De Santi & Sullivan, 1985; Dewitz & Dewitz, 2003; Duffelmeyer & Duffelmeyer, 1987; Keenan & Meenan, 2014; Keenan, Betjemann, & Olson, 2008; Nilsson, 2008; Taylor, 1983; Trezek & Mayer, 2015). Of those studies, eight explicitly described how IRIs relate to reading comprehension: IRIs measure student instructional levels in reading using a variety of methods (e.g., questions, oral retelling, rating scales). IRIs were often scored by total items correct and/or rubric rating scales.

Current evidence in the peer-reviewed literature does not support IRIs for screening and progress monitoring reading comprehension in K-12 settings. The lack of psychometric evidence is often cited as a critique of IRIs (Nilsson, 2008; Spector, 2012). Indeed, IRIs have demonstrated poor reliability evidence (Keenan et al., 2008) but variable-to-adequate intra-rater (De Santi & Sullivan, 1984) and alternate form (De Santi & Sullivan, 1985) reliability above the .80 standards (NCRTI, 2012). Of note, the reliability coefficient from Keenan et al. (2008) should be interpreted with caution given that it was calculated using unconventional standards. Overall, IRIs may not be a reliable measure or strongly correlated with criterion measures of reading despite their wide use and appeal.

Content-related validity evidence and usability evidence in the literature also did not fully support IRIs for screening. Elements of the IRI were found to be excessive or misleading in multiple cases. For example, Duffelmeyer and Duffelmeyer (1987) found that three informal reading inventories incorrectly classified main idea questions. Likewise, Clark et al. (2014) found that the student responses to prior knowledge questions on the QRI-4 did not relate to subsequent performance on relevant comprehension questions. This suggests that there are elements in IRIs that might seem appealing but lack evidence for their use. This would be problematic for screening because it would (a) increase time, (b) doesn't provide useful information, and (c) may not measure comprehension as described. In relation to progress monitoring, certain scores may not be aligned to the target domain or trait of reading comprehension (Kane, 2006).

Unlike the cloze, IRIs have been used with larger samples ranging from a single student to 995 (Mdn = 21.5; k = 8) spanning grades 3 through 11 (k = 5), and Trezek and Mayer (2015) investigated the utility of the IRI with deaf and hard of hearing populations. Despite this, the reported diversity of the sample was equally scarce as the research on cloze, in which students of color were represented 11% of the sample in a single study (Keenan & Meenan, 2014), and ELL-status was an exclusion criteria in two studies (Keenan et al, 2008; Keenan & Meenan, 2014).

Finally, Taylor (1983) found that 10% of the observed variation in scores on an IRI retell task was attributable to the interaction between teacher beliefs about English dialects and the student's speaking style. In short, teachers with unfavorable views of

diverse dialects rated students with different speaking styles lower on oral reading and retell tasks. This suggests differential rater behavior based on beliefs and student characteristics. If used for screening, students with different dialects that are tested by teachers with biases may be scored as further behind than their peers. This could result in differences in who is identified to receive intervention as well as influence group composition. In relation to progress monitoring, it could also impact the observed rate of improvement. As such, further research is needed.

The published peer-reviewed literature research (k = 11) appears to have emerging evidence regarding the IRI; however, those findings do not support the use of the IRI in K-12 settings as the sole tool for screening and decision-making. If IRIs are widely used in school settings, their validity evidence should be refined.

Maze unified validity evidence. A total of 19 studies provided validity evidence on the IUA of maze for students in K-12 (Brown-Chidsey, Davis, & Maya, 2003; Brown-Chidsey, Johnson, & Fernstrom, 2005; Espin & Foegen, 1996; Fore III, Boon, Burke, & Martin, 2009; Fuchs & Fuchs, 1992; Hale, Hawkins et al., 2011; Hale, Henning et al., 2011; Hale, Skinner, Wilhoit, Ciancio, & Morrow, 2012; Johnson, Semmelroth, Allison, & Fritsch, 2013; Marcotte & Hintze, 2009; Mccane-Bowling, Strait, Guess, Wiedo, & Muncie, 2014; McMaster, Wayman, & Cao, 2006; Muijselaar, Kendeou, De Jong, & Van den Broek, 2017; Price, Meisinger, Louwerse, & D'Mello, 2012; Reed, Vaughn, & Petscher, 2012; Speece et al., 2010; Stevenson, Reed, & Tighe, 2016; Tindal & Parker, 1989; Tolar et al., 2012). Of those studies, four explicitly described how maze relates to reading comprehension: maze is a modified version of the cloze task and measures the

accuracy and speed of sentence completion using multiple choice items. Broadly, maze measures sentence comprehension, vocabulary knowledge, syntactic understanding, and the ability to apply comprehension strategies during reading to build a mental representation of the text.

Overall, the peer-reviewed literature on maze provides the most information supporting its IUA in K-12 settings. Samples ranged in size and included diverse populations. Studies reported a median of 104.5 (Range 21-4215; *k* = 19) students were recruited from eight classrooms (range 2-24; *k* = 7) across one school (Range 1-15; *k* = 15) primarily in suburban areas (*k* = 5) in the midwest (*k* = 6) and southeast (*k* = 5) United States. Students of color represented 23% of the sample (Range 0-86%; *k* =14). The students were in grades 1 to 12 (*k* = 18). The median proportion of ELL students and students receiving special education services was 0% (Range 0-100%; *k* = 7), and 7% (Range 0-100%; *k* = 11), respectively.

In addition, the majority of maze tasks used across studies were commercially AIMSweb probes (*k* = 8), while a smaller number were developed by authors (*k* = 5). Psychometric evidence supports the maze for screening and progress monitoring in K-12. Test-retest reliability for maze met NCRTI (2012) standards when the same passage was used (r = 0.86) but not for different passages (r = 0.74; Tolar et al., 2012). Alternate form reliability was similarly variable in meeting standards (Johnson et al, 2013; McMaster et al., 2006). Internal consistency was not reported. Interscorer agreement ranged from 90% to 100% (*k* = 6). Criterion-related validity evidence (*k* = 4) suggested strong relationships between maze and the WJPC, Woodcock-Johnson Tests of Achievement,

Third Edition (WJ-ACH), WJ-ACH Broad Reading, WJ-ACH Word Attack, and Test of Emerging Academic English. There was also evidence of excellent classification accuracy and predictive-related validity evidence of maze to reading risk when paired with CBM spelling and teacher rating of reading problems (AUC = 0.92; R^2 = 0.49; Speece et al, 2010). Evidence supporting the use of maze in the schools suggests that it has high acceptability when compared to cloze, retells, and written retells (Fuchs & Fuchs, 1992).

The research on maze also investigated factors related to CIV. Maze tasks often cover general narrative or informational topics and are not linked to specific content (Johnson et al., 2009). In addition, students receiving special education services performed differently on maze than students who did not receive special education, and students who receive free and reduced lunch (FRL) appear to score similar to those who do not receive FRL (Stevenson et al., 2016). As such, it appears that maze scores do not systematically differ based on unrelated traits (i.e., FRL, SPED status, content knowledge).

Based on usability, reduction of CIV, and psychometric properties, the evidence in the peer-reviewed literature supports maze for screening reading comprehension. In contrast, the content validity evidence for maze was limited. No studies analyzed the content of maze tasks. A content analysis could be done with the answer choices. There are various methods for developing distractor items (e.g., the far and near method; Brown-Chidsey et al., 2003); however, current research has not examined the quality or the effects (i.e., different difficulties) of those item writing conventions. Overall, findings

are similar to cloze. Maze has potential for alternative deletion methods; however, these also have not been explored. The maze task in the literature was often defined as an alternative to CBM-R. To further illustrate the connection, in one factor analysis, maze was grouped under a reading fluency factor (Muijselaar et al., 2017).

The published peer-reviewed literature (k = 19) appears to support the maze for screening reading comprehension in K-12 settings. In conclusion, maze has exceptional psychometric and usability evidence, yet the direct connections to reading comprehension are sometimes unclear and could be improved.

Multiple choice reading comprehension unified validity. A total of six studies provided validity evidence on the IUA of multiple choice reading comprehension (MCRC) scores for students in K-12 (Hale, Henning et al., 2011; Hale, Hawkins et al., 2012; Mccane-Bowling et al., 2014; Neddenriep, Hale, Skinner, Hawkins, & Winn, 2007; Skinner et al., 2009; Walczyk, 1990). All studies used some variation of Timed Reading Series (Spargo, 1989). This was due to the inclusion procedures, which required assessments to have reproducible methods or a specific developer. Two studies explicitly described how MCRC relates to reading comprehension: MCRC uses factual and inferential questions to measure inferences made during reading.

The peer reviewed literature does not provide sufficient information to support the IUA of MCRC assessments, specifically Spargo (1989), for screening or progress monitoring. As stated earlier, a score must be reliable to support valid interpretation and use (Kane, 2013). Unfortunately, it is currently unclear if scores on MCRC are reliable. Studies did not report test-retest reliability, and relationship between alternate forms did

not meet standards (Walcyzk, 1990). However, interscorer agreement ranged from 97% to 100% (*k* = 3). Criterion-related (*k* = 3) validity evidence also suggested relationships between the MCRC and WJPC, WJ III-ACH, Iowa Reading Comprehension subtest. Thus, MCRC scores appear to share systematic variation with reading comprehension; however, it is unclear if these scores are consistent over multiple testing periods or forms.

Generally, MCRC appears to be a quick measure of reading comprehension following an oral or silent reading task. It demonstrates evidence of predictive validity evidence when used in combination with maze and CBM-R (Hale, Henning et al., 2011; Neddenriep et al., 2007). However, the importance of these findings are contingent on adequate reliability estimates. In addition, the representativeness of the samples is unclear. Overall, studies reported that a median of 36.5 (range 22-98; *k* = 8) students were recruited from two (range 1-4; *k* = 3) classrooms, across one school (range 1-2; *k* = 5) in the southeast United States (*k* = 3). Students were in grades 3 through 11(*k* = 7). Students of color represented 12% (range 8%-56%; *k* = 5) of the sample. No data were reported regarding ELL students, and no students received special education services (*k* = 4).

The published peer-reviewed literature (*k* = 6) currently lacks sufficient evidence to support the IUA of MCRC as a reliable estimate of reading comprehension in K-12 settings. Additional research is needed on MCRC to establish consistency of scores (i.e., test-retest and alternate form reliability).

Retell unified validity. A total of ten articles (11 studies) provided validity evidence on the IUA of retell scores for students in K-12 (Bernfeld, Morrison, Sudweeks,

& Wilcox, 2013; Carlisle, 1999; Fore III et al., 2009; Fuchs & Fuchs, 1992; Fuchs et al., 1988; Hansen, 1978; Marcotte & Hintze, 2009; Reed et al., 2012; Shapiro, Fritschmann, Thomas, Hughes, & Mcdougal, 2014; Tindal & Parker, 1989). Of those studies, six explicitly described how retell relates to reading comprehension: retell is a procedure in which students are tested on their ability to read a text, form a mental representation of it, and communicate it (written or orally) in an organized way that is both relevant and highlights key content.

Regarding the sample, the retell was used with moderately sized samples that diversity in the racial and ethnic, linguistic, and education status of its participants. These studies were conducted across elementary, middle, and high school (k = 7). Studies reported that a median of 70 students (Range 30-311; k = 9) were recruited from five classrooms (Range 3-6; k = 3), across two schools (Range 1-7; k = 5) in suburban areas (k = 2). Students of color represented 32% (Range 14-86%; k = 5) of the sample. The proportion of ELL students was 24% (Range 0-30%; k = 3), and 30% of students received special education services (Range 0-100%; k = 9).

Despite range of participants, the retell procedure was not supported for screening or progress monitoring in K-12 for reading comprehension. Similar to MCRC, reliability evidence must be investigated. Test-retest and alternate-form reliability and internal consistency were not reported. This is problematic given that a common difficulty of the retell task is scoring student verbal responses. Bernfeld et al. (2013) found large standard deviation differences between scoring methods (Cohen's d range: 3.83 - 4.12) when comparing real time and audio scoring.

The sources of systematic error in the retell procedure appear to be partially a result of scoring. Developing scoring procedures with higher agreement is needed. Even tasks that merely required counting the total words resulted in highly discrepant scores on average (Bernfeld et al., 2012). Some authors went through extensive procedures and decision-making processes to code similarly on a set of passages using a four-item rubric (Reed et al., 2013). In addition to scoring, retells can be timely, taking 1 to 30 minutes to administer per student (k = 5). In addition, retell can be easily administered following an oral reading task and would take one-minute to administer per student. Paired with the inconsistency of scores, the measure may serve best as a supplemental measure of reading comprehension.

Similar to the MCRC tasks, retell does share systematic variation with measures of reading comprehension. Specifically, criterion-related validity evidence suggests a relationship between retell and the Stanford Achievement Test – Reading Comprehension (Fuchs et al., 1988). Evidence suggests low acceptability of both the oral and written retell procedure, compared to the maze. Specifically, teachers were concerned that the total words scoring methods was mainly a measure of oral production (Fuchs & Fuchs, 1992).

Ultimately, the published peer-reviewed literature research (k = 11) lacks the evidence to support the IUA of retell in K-12 for screening. There is preliminary evidence; however, in the context of the other available measures, there is less support until additional improvements can be made to the assessment and once more evidence is gathered, in refining administration and scoring procedures.

Sentence verification technique unified validity. A total of 11 studies provided validity evidence on the IUA of sentence verification technique (SVT) scores for students in K-12 (Carlisle, 1989a; Carlisle, 1989b; Carlisle, 1991; Carlisle, 1999; Marcotte & Hintze, 2009; Rasool & Royer, 1986; Royer & Carlo, 1991; Royer, Hastings, & Hook, 1979; Walczyk & Royer, 1989; Walczyk, 1990). Each study explicitly described how SVT relates to reading comprehension: SVT measures if readers establish and maintain a mental representation of the text. Students must determine if a series of exact, incorrect, or paraphrased sentences accurately represent information from the passage.

In regard to sample, the SVT was used with a range of students. Studies reported that a median of 48 students (Range 2-315; $k = 10$) were recruited from three classrooms (Range 3-6; $k = 3$), across two schools (Range 1-2; $k = 3$) in the northeast United States ($k = 4$). Students of color represented a median of 63% of the sample (Range 33-50; $k = 4$). The students were in grades 4 through 8 ($k = 4$). The proportion of ELL students was a median of 46% (Range 30-61%; $k = 2$) and a median of 29% (Range 20-30%; $k = 3$) of students received special education services.

The SVT has emerging evidence to support its IUA for screening reading comprehension in K-12. The psychometric evidence could be improved. Internal consistency met standards for reliability in grades 4-6 ($\alpha = 0.81$-0.84) but not in third grade ($\alpha = 0.76$; Royer & Carlo, 1991). Interscorer agreement was not reported. Criterion-related validity evidence did not meet standards (Marcotte & Hintze, 2009; Royer, Hastings, & Hook, 1979). Although internal consistency was adequate (Royer & Carlo, 1991), criterion-related validity evidence and classification accuracy were not

sufficient. However, it is noteworthy that SVT showed evidence of classification accuracy with the Profiles in Listening and Reading (PILAR) and Gates-MacGinitie (Gates) tests, when SVT was included in a battery of tests (i.e., reading and listening SVTs and with word identification assessments; Carlisle, 1989a).

In relation to factors related to CIV, studies were primarily from the 1980s, and all studies were connected to the original research team. Furthermore, the population was primarily students in suburbs in the state of Massachusetts. Thus the current evidence may not generalize to current-day school-aged populations.

Of note, Royer and Carlo (1991) examined the utility of the SVT in identifying levels of English proficiency. It was found to distinguish between students in mainstream classrooms and students with varying years of English instruction. Unfortunately, the psychometric evidence such as reliability and predictive utility were not disaggregated, suggesting that more research could be conducted in this area given that score consistency is relatively unknown.

The published peer-reviewed literature (k = 11) demonstrates promising evidence of the IUA of SVT in K-12 for screening and progress monitoring. Future research could expand the availability of materials, replicate and extend key findings on reliability and update criterion and diagnostic validity evidence of SVT to criterion measures, and extend the use of the measure to different grade levels and populations.

Think-aloud unified validity. Two studies provided validity evidence on the IUA of think-aloud results for students in K-12 (Meyers, 1988; Meyers & Lytle, 1986). Both studies explicitly described how think-alouds relate to reading comprehension:

think-alouds have the reader communicate their thoughts as they read a text in order to observe online comprehension and cognitive strategies. Think-alouds were delivered and scored by a school psychologist. The student's cognitive strategies (i.e., moves) were coded for themes and errors. The authors did not report administration time, psychometrics, or factors relevant to CIV.

Overall, both studies recruited one 4th grade student from an unknown number of classrooms and schools that were in suburban areas. Details on race, special education status, and English language proficiency were not reported.

The current evidence on think-alouds in K-12 is too scarce to make any decision regarding screening. Across all domains, there was little evidence of methodological characteristics, content, usability, factors related to CIV, and psychometric validity evidence. These findings are surprising given that think-alouds are a commonly used approach to understand the metacognitive strategies of adults during reading (e.g., Meyers & Lytle, 1986). The authors discussed the potential for think-alouds to be used as diagnostic assessments; however, only broad and anecdotal statements were provided about the effectiveness of those decisions (Meyers, 1988; Meyers & Lytle, 1986).

The published peer-reviewed research ($k = 2$) lacks the evidence to support the use of the think-aloud in K-12 settings. Reliability and usability evidence must first be attained using a replicable procedure to understand the utility of think-alouds.

Summary. There is variability in the validity evidence supporting the IUA of reading comprehension measures in K-12 settings. Specifically, the variability in evidence is particular to the type of measure utilized (e.g., MCRC, retell). In the context

of schools, measures are used for a variety of purposes. Specifically, they are commonly used in relation to intervention and instruction (Salvia et al., 2007).

As stated previously, students are struggling in reading, and the purpose of reading is comprehension (NRP, 2000). Thus, it is imperative that education researchers and practitioners find measures to identify students with reading problems and monitor the effectiveness of the supports provided (e.g., intervention and instruction). Not surprisingly, the majority of the purposes of assessment relate to the selection, monitoring, and evaluation of instructional and intervention practices (Salvia et al., 2007). Thus, the identification of measures with validity evidence for use in schools as well as the continued improvement of additional measures is paramount to the continued assessment and intervention process in reading comprehension.

Impetus for the current research

In summary, the review of the reading comprehension assessment literature returned information relevant to the validity evidence supporting the IUA of several reading comprehension measures: cloze, IRI, maze, MCRC, retell, SVT, and think-alouds. Each assessment technique had its own strengths and weaknesses in relation to the types of evidence present in the peer reviewed literature. Despite the differences in validity evidence, each of the measures have been used for assessing reading comprehension performance in students. Many purposes of assessments in the school settings are related to instructional and intervention practices (Salvia et al., 2007). Very few meta-analyses (i.e., Hebert et al., 2013) published in the peer-reviewed literature have investigated the relationships between variability in intervention outcomes based on

the particular measure of reading comprehension used beyond the standardization of the task. However, evidence suggests that different measures of reading comprehension measure reading comprehension differently (Collins et al., 2018; Garcia & Cain, 2014; Kendeou et al., 2014; O'Reilly et al., 2014). Thus, additional research is needed.

Based on the current research, a gap in the current peer-reviewed literature on reading comprehension is evident. There is emerging evidence to support the IUA of reading comprehension measures in elementary school, and one of the primary purposes of their use is for intervention selection, evaluation, or monitoring. However, it is unclear if certain measure and intervention combinations lead to different results. As such, the purpose of the present dissertation was to investigate if intervention effectiveness varied based on the measure's response format.

Given the large amount of individual intervention studies involving reading comprehension, this study specifically used meta-analysis to understand and explore possible differences in intervention effectiveness. The goals of the research were to highlight differences in assessment and intervention pairings and address a gap in the literature that is relevant to both research and practice. The research was also conducted to support the identification of reading comprehension measures that differentially measure particular interventions or practices.

Beyond the simple classification of a measure as a standardized criterion-referenced measure or experimentally designed, it is still unclear if intervention effectiveness is impacted by the reading comprehension measure used. The current study

aimed to address the current gap in the literature through meta-analysis. Specifically, three questions were examined:

RQ 1: In standardized mean difference, as measured through Hedge's *g*, what is the observed aggregate treatment effect for all reading comprehension interventions delivered to students in elementary school at posttest compared to control?

RQ 2: In standardized mean difference, as measured through Hedge's *g*, what is the observed aggregate treatment effect for all reading comprehension interventions delivered to students in elementary school from pretest to posttest?

RQ 3: Do observed treatment effects differ based on the specific interventions and measures?

Chapter 3: Method

Search Procedures

The comprehensive search used (a) online databases, (b) unpublished grey literature, (c) reference lists from relevant meta-analyses, and (d) key informants, which is consistent with published recommendations for meta-analytic methods (Card, 2012; Cooper et al., 2009; Harwell, 2008). To begin, a systematic search of scholarly literature and dissertations was conducted using several databases across disciplines: Academic Search Premiere, Educational Resources Information Center (ERIC), Linguistics and Language Behavior Abstracts (LLBA), ProQuest Dissertations and Theses (PDAT), and PsychInfo. Search terms were identified using relevant terms from 39 meta-analyses on reading comprehension. Searches in the literature were typically broad and relied on a few terms. As such, two terms were selected, *reading comprehension* and *intervention*. These terms were applied to each database utilizing Boolean expressions. Thus, databases were used to search for articles that included the terms *reading comprehension* AND *intervention* in either their title or abstracts.

In addition to doctoral dissertations, unpublished grey literature was obtained by contacting authors (Card, 2012; Field & Gillett, 2010). Authors were contacted if they (a) appeared as first author in two or more included studies, or (b) were first author on a relevant meta-analysis investigating outcomes of reading comprehension interventions and made efforts to include grey literature in their analyses. This was done because the meta-analysis researchers may have had data or knew of additional researchers with relevant data to share. The list of authors is reported in Appendix A. Each informant was

contacted using the same email request. If there was no response after two weeks, two follow-up emails were sent, each separated by two weeks. Information regarding this process was documented through email and Microsoft excel; however, ten authors were identified. Nine had relevant contact information and were emailed for unpublished data. Seven responded, sharing that they had no unpublished data to provide for various reasons, while the remaining two authors did not reply.

Inclusion and Exclusion Criteria

Inclusion criteria were developed based on the research questions and prior research discussed in Chapter 2. As a result, a total of eight gates were established. Criteria for each was necessary to be included in the current meta-analysis. To begin, duplicates were removed. In addition, it was a requirement that studies were published in 2000 or later in order to ensure the most recent research represented the results (Swanson et al, 2017). Some evidence (Lemons, Fuchs, Gilbert, and Fuchs, 2014) suggests that the business as usual (BAU) condition (e.g., tier 1 instruction) has improved over time, which has resulted in the attenuation of observed differences in performance between treatment and control groups. As such, in order to control for this variation, study year was noted and the range was restricted for relevance. This procedure mirrors Swanson et al.'s (2017) procedures. Then, the following sequential methods were used to identify studies eligible for inclusion in the meta-analysis. Studies were included in the analyses if they:

1. (Gate 1) were written in English, conducted in elementary school settings (K-5 or K-6 or disaggregated by grade level; e.g., Suggate, 2016; Swanson, 2017) in the United States, related to reading, and a primary source;
2. (Gate 2) implemented an intervention (e.g., Suggate, 2016; Swanson, 1999);
3. (Gate 3) implemented a reading comprehension intervention as opposed to fluency, vocabulary, phonemic awareness, and phonics interventions (Suggate, 2016);
4. (Gate 4) collected outcome data on reading comprehension that related to (a) a relevant experimental measure (i.e., cloze, IRI, maze, MCRC, retell, SVT, think-aloud) or (b) a standardized, norm-referenced and commercially available measure (e.g., reading comprehension subtest of an achievement test);
5. (Gate 5) used an experimental or quasi-experimental group design;
6. (Gate 6) collected data at pretest and posttest on reading comprehension using a relevant measure of reading comprehension for both the intervention and control; and
7. (Gate 7) provided sufficient data to calculate effect sizes (Card, 2012).

Studies that did not meet the criteria for a particular gate were excluded, and that information was logged in a spreadsheet. Of particular note, Gates 2 and 3 were separated to improve the specificity of the process and to identify other studies for future reviews (e.g., Gate 3 would be useful for future reviews pertaining to reading interventions such as fluency or vocabulary, while Gates 2 and 3 combined would be much broader).

Missing Data. A commonly reported issue in conducting meta-analysis is the magnitude of missing and unreported data within and across studies (e.g., Berkeley et al., 2015; Li, 2014; Scamacca et al., 2015). Thus, it was expected that some studies would not include sufficient information to pass though all seven gates. In order to include a representative sample of studies from this body of research, the first author contacted first authors of studies that met all criteria except Gate 7. Authors were contacted using a modified version of the emails presented in Appendix A. Those materials are provided in Appendix B. When no response or data were provided after a maximum of three possible emails, the study was excluded.

Sample

The search procedures returned a total of 4,199 articles. Duplicates were removed, resulting in a total of 2,517 studies. The majority of these studies did not meet inclusion criteria: 1,281 studies were excluded at Gate 1, 413 at Gate 2, 394 at Gate 3, 78 at Gate 4, 187 at Gate 5, 14 at Gate 6, and 38 at Gate 7. A total of 63 studies were included in this review from these procedures. In addition, citation searches were conducted across the meta-analyses of reading comprehension reviewed in chapter 2. A total of 69 articles were identified through references. Of those articles, 20 were duplicates. Inclusion criteria were applied to the remainder. At Gate 1 30 studies were excluded; at Gate 2 three studies were excluded; at Gate 3 seven studies were excluded; at Gate 4 four studies were excluded; no studies were excluded at Gate 5; one study was excluded in Gate 6; and one study was excluded at Gate 7. As a result, 46 additional studies were excluded, and a total of three studies were included from this method.

Ultimately, 66 studies were included in the meta-analysis. Figure 1 uses the PRISMA model to visually represent this process (Moher, Liberati, & Altman, 2009). Those studies were then systematically coded by the first author, and 20% of the total studies were coded by a trained independent rater who was a graduate student in educational psychology. This was calculated by the percentage of total agreements between coders, and a value of 90% or more appears to denote high agreement (Harwell, 2008). Inter-coder agreement was 92.5%.

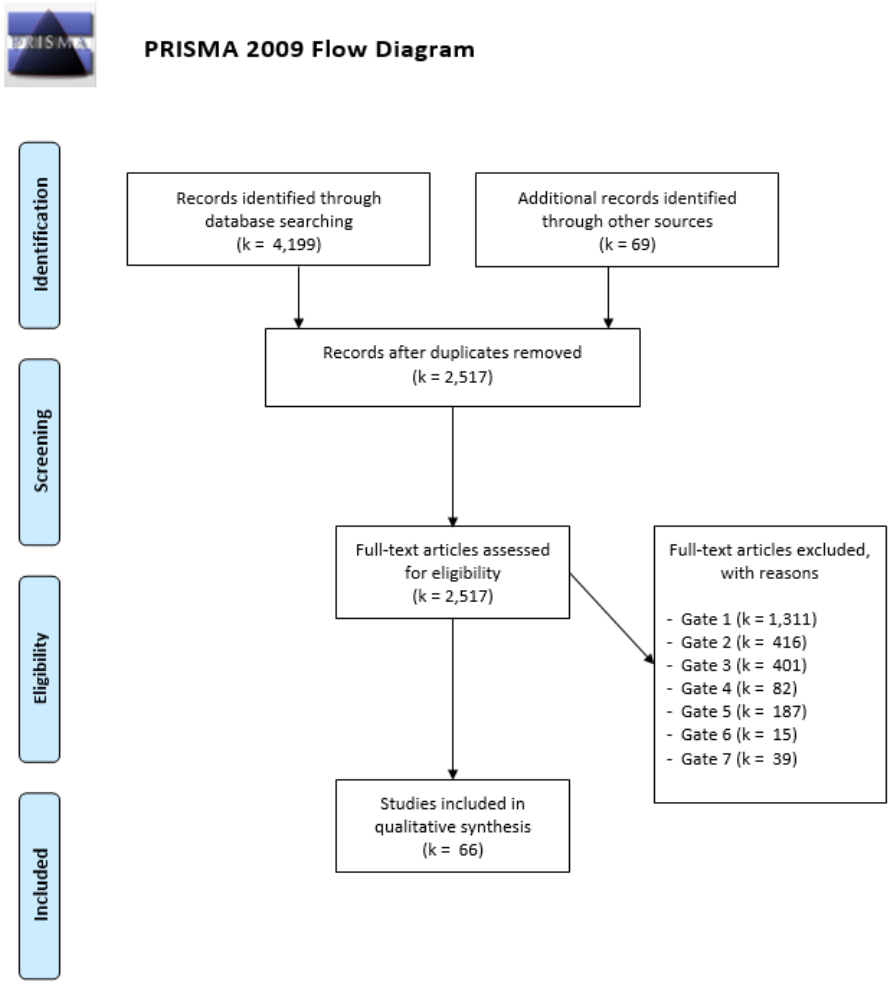

Figure 1. Flow diagram of study search and inclusion procedures. *From:* Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Coding

Each study was coded for a variety of characteristics relevant to the (a) basic study information, (b) individual assessment info for each study, (c) study design, (d) intervention characteristics, (e) reader characteristics, (f) effect size information, and (g) quality indicators (i.e., the Council for Exceptional Children [CEC; 2014] standards). Each of these are defined in detail below. The codes were developed based on the research questions, previous meta-analyses on reading comprehension, NRP (2000), and coding procedures used in previous meta-analyses in reading comprehension.

An initial codebook was developed based on these criteria and refined through iteratively coding one article. Additional coders were trained by practicing on that article. This procedure was based on the approach used by Murphy et al. (2009). In addition, an excel sheet was created with dropdown options to minimize disagreements. The NRP (2000) procedures required that at least 10% of the included studies were coded by an additional coder. This study followed those procedures but coded 20% of the included studies, as is standard practice.

(a) Basic study information. The purpose of the study information section was to code descriptive information at the broad study level. This included the following: creating a study identification variable and recording the date coded, authors, year, title,

publication type, citation, citation: short form (e.g., Author et al, 2008), program affiliation of the first author, a narrative summary, and the initials of the coder. Each code represented in the section is a reflection of Collin's et al. (2018) coding manual, the NRP (2000) coding procedures, and recommendations provided by committee member(s).

(b) Assessment. Assessment characteristics were coded based on Collin's et al. (2018) coding procedures; they investigated sources of variance for reading comprehension assessments for students with and without comprehension difficulties. Therefore, their coding manual, in part, focused on features of assessments. Those elements were used to create codes for assessment characteristics within the current study. In addition, NRP (2000) was used; specifically, the latency between the pretest and posttest. Thus, the following codes were used for assessment characteristics: assessment identification number (created), name of RC measure, standardization, measure description, response format, score reliability of reading measures, reverse scored, administration size, measurement ceiling or floor effects (treatment and control), fidelity of assessment administration, student scores reporting format, administration of measure, type of text/genre, timed vs untimed measures, background knowledge assessed, student reading (i.e., whether the students read the text, orally, silently, both, or in a different way), if reading assistance provided, passage viewing (i.e., whether the passage was able to be reread during the assessment), passage length, passage difficulty, administration of questions, probed/unprobed recalls, oral/written recalls, and pre-post latency (defined by number of weeks between the pretest and posttest administration).

(c) Study design. Study design was coded using the procedures in the NRP (2000) report. These included, whether the study was a true experiment or a quasi-experiment, the level of randomization, if matching was used in the randomization process, whether matching or statistical control was used to address nonequivalence issues in quasi-experiments, description of how the sample was attained, any attrition, description of the control conditions, and description of the treatment conditions.

(d) Intervention characteristics. Intervention characteristics were coded based on a variety of sources. To begin, the NRP (2000) report was used to gather a basic understanding of the independent variable and its implementation. Reading comprehension intervention categories were selected based on the classifications used in several meta-analyses and in Joseph's (2015) review of reading comprehension interventions. These specifically included strategy instruction, text structure instruction, improving background knowledge, self-monitoring, inference instruction, and graphic organizers. Interventions that included more than one component were classified as multicomponent. The diversity of reading comprehension interventions implemented in schools is evident based on the number of intervention meta-analyses that currently exist. Thus, any additional interventions that were not best captured by the aforementioned categories were coded as, other.

Generally, these were the codes for this category: if the universal curriculum was described, description of the curriculum, intervention setting, whether the intervention was delivered in an implicit or explicit format, minutes per session, sessions per week, number of weeks, whether fidelity of treatment was assessed, number of implementers,

interventionists per student, characteristics of the implementer, length of training, source of training, how implementers were assigned to groups, whether continued consultation was provided, whether the intervention included a writing component, graphic organizer, signal words (e.g., "*infer*"), or vocabulary component, the type of intervention (i.e., strategy instruction, text structure instruction, improving background knowledge, self-monitoring, inference instruction, graphic organizers, multicomponent, or other), and the intervention's name.

(e) Reader characteristics. Reader (i.e., participant) characteristics were coded based on common conventions and topics of previous meta-analyses (i.e., special education, reading disability, and reading risk status [Suggate, 2016]; English language learner status and native language [Spencer & Wagner, 2017]; race and ethnicity; gender; grade; indicators of socioeconomic status such free and reduced lunch status and mother education). These factors were primarily accounted for in Collins' et al. (2018) coding manual.

As such, the following were used as codes for this category: grade range, whether students with reading disabilities were identified, race and ethnicity (i.e., white, black, Asian, Hispanic, other), whether students received subsidized lunch, state or region of the population, urbanicity, number of schools, number of classrooms, whether the population included any groups of exceptional learning students (i.e., learning disability, reading disability, deaf/hard of hearing, autism, or other), English language learners or limited English proficient students, and whether the sample was restricted to include or exclude

certain populations (e.g., only students with emotional and behavioral disabilities, adequate decoding).

(f) Effect size information. Codes for effect size information were determined based on Collins et al. (2018) and NRP (2000). Thus, codes for effect sizes included individual effect size identification numbers; the mean, standard deviation, and sample size for both the treatment and control groups at pretest and posttest; statistics other than means and standard deviations used to calculate the effect size (e.g., *F*-statistics, Cohen's *d*; Card, 2012); the effect size for student growth from pretest to posttest; the effect size for group differences at posttest.

(g) Study quality. Study quality was coded using NRP (2000) to describe the methodological approaches used. Study quality was also coded using criteria from CEC (2014). These standards were used given their relation to education research and assessing quality for both group and single case design studies. Specifically, (a) the extent to which the control condition had access to the intervention, (b) whether the overall attrition was below a 30% criterion, and (c) whether the study reported outcomes for all target measures and not only those of statistical significance (CEC, 2014). Scores were given to studies for having all of the information or no disconfirming information (2), at least one part of the information (1), or none of the information (0). This was done in order to produce a more systematic and quantified method for interpreting the standards. As a result, greater quality scores denote higher study quality.

Analyses

Random effects. Random-effects models are used when study results vary across the population; specifically, a random effects model assumes a homogenous effect size is not shared across the population (Card, 2012). Thus, there is no one fixed effect size. Instead, effect sizes vary across studies based on the populations that were sampled (e.g., intervention type, reading proficiency [Collins et al., 2018; Garcia & Cain, 2011]). As such, the random-effects model assumes that there are multiple, true effect sizes that vary by population. As a result, random effects model assume effect sizes for each study vary due to between and within-study sampling error. The assumption allows the researcher to make inferences about studies and study characteristics beyond the studies represented in the sample since it recognizes that the included studies are only a sample of the population.

In contrast, a fixed-effects model is used when one true effect size for reading comprehension interventions is assumed, and differences between effect sizes are attributed to within-study sampling error. Unlike the random-effects approach, the fixed effects does not allow inferences beyond the sample of included studies. This is because the mean effect size is fixed as a function of the current sample of data.

The current meta-analysis adopted a random-effects model due to the heterogenous population of studies that exist, measuring different student groups, types of readers, interventions, and the broad construct of reading comprehension, and conditioning on these study characteristics is insufficient to fully explain the heterogeneity of effect sizes since there is a distribution of true effect sizes that create the

universe of effect size estimates that could be sampled (Bornstein, Hedges, Higgins, & Rothstein, 2011).

Effect size calculation. The unit of analysis for this study was at the effect size level. Effect sizes are nested within studies. In the context of this meta-analysis, an effect size represented the magnitude of a treatment effect. The treatment effect was the observed difference score. This was examined as both the difference between (a) the pretest to posttest score for the treatment group as well as (b) the difference between the treatment and control groups at posttest. Hedge's *g* effect sizes were calculated for each included study. Hedge's *g* was calculated using the following formula (Borenstein et al., 2009):

$$g = \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right) * \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}},$$

where the first part of the equation represents a correction for small sample sizes:

$$\left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right),$$

and the latter half of the equation represents the Cohen's *d* effect size estimate:

$$\frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}},$$

where Cohen's *d* is equal to the difference between means divided by the pooled standard deviation, and where the pooled standard deviation is the weighted standard deviation. The equation may be simplified as

$$g = \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right) * \frac{M_1 - M_2}{S_{pooled}}.$$

Hedge's *g* provides a correction for small sample sizes. Thus, no additional corrections for sample size were included. Of note, for the first research question, M_1 represented the treatment group while M_2 represented the control group. For the second research question, M_1 represented the treatment group at posttest while M_2 represented the treatment group at pretest.

Heterogeneity and confidence intervals. Several other noteworthy metrics were used to interpret findings. *Q* is a measure of heterogeneity between effect sizes (Card, 2012). It is calculated by

$$Q_j = \sum w_j g_j^2 - \frac{\left(\sum w_j g_j\right)^2}{\sum w_j},$$

where, w_i is the weight of the *j*th study, and g_j is the effect size estimate of the *j*th study. A significant *Q* statistic suggests heterogeneity between effect sizes. Of note, *Q* follows a chi-square distribution with *m*-1 degrees of freedom, where *m* is the number of effect sizes. Thus, its significance may be interpreted using a chi-square table (Card, 2012).

Although the studies all implemented reading comprehension interventions and were aggregated for those reasons, they varied based on a number of elements (e.g., type of outcome measure, intervention duration, age group). Thus, it was important to measure the extent of heterogeneity between studies that is not due to sampling, which *Q* does. In addition to *Q*, I^2 provided an estimate of the percentage of heterogeneity between study effect sizes compared to the total variability in effect sizes (Card, 2012). I^2 was calculated by,

$$I^2 = \frac{Q - m - 1}{Q} * 100,$$

when $Q > m–1$ and I^2 is zero when $Q < m–1$. Card (2012) reported that I^2 may be interpreted as small (25%), medium (50%), or large (75%) heterogeneity, or it may be interpreted as homogenous (0%). Together, Q denoted whether there was heterogeneity, and I^2 described how much heterogeneity.

In addition, confidence intervals were used to estimate possible ranges of a true treatment effect. When a confidence interval included zero, it was indicative of a null effect, either because the true score could be zero or because the standard error was too large.

Analytic procedures. First, two overall effect sizes for the effect of reading comprehension interventions on reading comprehension assessment outcomes were calculated. These involved the standardized mean difference between (a) treatment and control at posttest and (b) the treatment group score from pretest to posttest.

The estimation of the overall effect size for the standardized mean difference between treatment and control at posttest was subjected to a sensitivity analysis. A sensitivity analysis was conducted to examine if findings varied between studies that were true experiments compared to all relevant studies (including those that utilized quasi-experiment designs). No differences were used to support the inclusion of all studies in the analysis for research question 1. The overall effect size research question 2 was calculated using all relevant individual study effect size estimates, and a similar sensitivity analysis as conducted in research question 1 was used.

The overall effect sizes calculated for both research questions incorporated each dependent variable of reading comprehension that was measured at pretest and posttest or between groups. In order to help control for bias in the overall effect sizes, an additional weight was used for studies with multiple effect size estimates (i.e., studies with multiple effect size estimates are reported as, m_j, where there are *m* effect sizes in study *j*). Specifically, the weight was shared evenly between multiple within-study effect sizes. Thus, if a study reported two relevant effect sizes (e.g., $m_j = 2$), both were included in the analysis, and their variances were multiplied by two. The original formula for weight was, $w = \frac{1}{\Delta}$, where *w* was the weight, and Δ was the conditional (sampling error) variance associated with the within-study effect size. The adjusted weight for the effect size involved multiplying the inverse variance by the total number of effect sizes (*m*) within that particular study (*j*). Thus, $w_{adjusted} = \frac{1}{\Delta_j * m_j}$.

Heterogeneity was examined within the overall effect sizes. When significant heterogeneity was observed, study-level variables were examined, particularly, to determine if effect sizes varied based on the measure and intervention used. In addition, other study characteristics relevant to (a) basic study information, (b) individual assessment info for each study, (c) study design, (d) intervention characteristics, (e) reader characteristics, and (f) quality were examined.

These variables were entered in blocks according to their category (e.g., assessment, quality) to determine if they explained significant variance in effect sizes. This was done using regression. Specifically, random-effects weighted least squares regression with maximum likelihood estimation (Raudenbush & Bryk, 2002) was used to

value sources of variance that contributed to potential differences between (a) students who receive reading comprehension interventions compared to control at posttest and (b) student growth from pretest to posttest. Variance was accounted for at the levels of effect size (level 1 within-study variance represented as study-specific sampling variance) and study (level 2 between-study variance represented as a random-effects variance component).

Level 1: $g_j = \delta_j + e_j$

Level 2: $\delta_j = \gamma_0 + \gamma_1 W_{1j} + \gamma_2 W_{2j} + \ldots + \gamma_s W_{sj} + u_j$

with the combined mixed model of: $g_j = \gamma_0 + \Sigma\gamma_s W_{sj} + u_j + e_j$

where, for level 1, g_j was the study estimated effect size for study j, so that g_j is normally distributed: $g_j \sim N(\gamma_0 + \Sigma\gamma_s W_{sj}, \tau + V_j)$ where $Var(g_j) = \tau + V_j = \Delta_j$. δ_j was the true effect size estimate and e_j was the sampling error associated with g_j as an estimate of δ_j, where $e_j \sim N(0, V_j)$. The within-study variance of the sampling error was estimated based on the sampling variance of g_j, to compute the weight for the WLS analyses, and then associated with each g_j in a variance-known model. The reciprocal of the variance (Δ_j) of each effect size (m) was the weight that produced an efficient estimator in the WLS model.

For level 2:

γ_0 was the mean effect size, adjusted mean given the coding of W_s.

γ_s was the regression coefficient associated with study characteristic W_s.

W_{sj} included each study-level variable s for study j.

u_j was the unique effect for each study where $u_j \sim N(0, \tau)$. This between-study variance, τ, was the maximum likelihood estimate of the random variance component to

complete the random-effects model, which was added to the sampling error variance to create a weight that minimizes the variance of the estimated mean effect size, such that $\text{Var}(g_j) = \tau + V_j = \Delta_j$.

Follow-up analyses were conducted when results indicated that effects may vary based on the outcome measure. Separate effect sizes for discrete dependent variables (e.g., maze, Woodcock Johnson-Passage Comprehension) were used to examine effects based on each measure.

Publication bias. Publication bias was assessed to examine if studies were excluded that should have been included. Publication bias has the potential to influence results, suggesting that the findings represented are not accurate or representative of the true effect of reading comprehension interventions. Due to the file drawer problem, it's possible that the observed outcomes in meta-analytic studies are inflated because nonsignificant findings are less likely to be published (Card, 2012).

Publication bias was assessed through a funnel plot, which plots effect sizes on the x axis and sample sizes on the y axis. Funnel plots were interpreted using the following rationale: a lack of publication bias was associated with symmetry in the plot while publication bias was associated with asymmetry.

Reporting of results. Results were reported in terms of summary tables that aggregated descriptive information across studies (e.g., median sample size, date range, total effect sizes, proportion of experimental to standardized measures). Findings from the meta-regressions were similarly reported in tables. Although other standards exist, effect sizes were interpreted using the conventional standards in reading comprehension

meta-analyses (e.g., Scammacca et al., 2015; Spencer & Wagner, 2018), which use Cohen's (1988) suggestion for the interpretation of effect size estimates: small (ES $\geq$ 0.20), medium (ES $\geq$ 0.50), and large (ES $\geq$ 0.80).

Software. Microsoft Excel and R were used to conduct this meta-analysis. Data were coded using an Excel spreadsheet. Those data were then loaded into R. The metafor package and the rma function were used to conduct the meta-analysis and meta-regressions. All categorical variables were entered as factors (i.e., dummy-coded).

Chapter 4: Results

Meta-analytic methods were used to explore the relationship between reading comprehension intervention effects and the measures used to quantify those effects. The results are organized to provide descriptive information for the sampled studies. This is followed by a brief description and tests of the analytic assumptions. Finally, there is one section to describe the results for each research question (RQ).

Descriptive Data

A total of 66 (*k*) studies were included in this meta-analysis, where the total number of studies is denoted by *k*. They represented a total of 220 (*m*) effect sizes, which is denoted as *m*. Studies were published from the years 2000 to 2019. There was a median of 30 students (M = 72.39; SD =131.34) in the treatment group across effect sizes, and there was a median of 25 students (M = 60.19; SD = 121.82) in the control conditions.

Table 1 contains descriptive information for relevant study-level design, quality, and reader characteristics. The variables represented in Table 1 exceed the total number of studies (i.e., k = 66) because two studies included in the meta-analysis (Allor & Mccathren, 2004; Amendum, Vernon-Feagans, & Ginsberg, 2011) contained two separate samples. As a result, there were 68 cells of information, meaning that total cells exceeded the number of studies (k = 66). Therefore, n_{cells} is used to represent the amount of relevant data associated with the variables, and the total amount of 68 relevant cells of data possible (i.e., 68) is denoted by N_{cells}. As such, Table 1 contains relevant study-level characteristics across k = 66 studies, with a total of N_{cells} = 68 of information.

Table 1

Study-level Descriptive Characteristics for Included Studies (N_{cells} = 68)

Variable	n_{cells}	*Median*	*M*	*SD*
Attrition	58	0	17.93	50.74
Design (True Experiments)	50			
ELL/LEP Students				
Missing	47	32.13	48.95	40.39
Level of Randomization				
Student	29			
Group	15			
School	7			
Other	0			
Missing	17			
Control Conditions				
Business as Usual	53			
Treated Control – SD	11			
Treated Control – DD	0			
Other	4			
Grade (Min)	68			
Grade (Max)	68			
Publication Year				
Since 2010	49			
Before 2010	19			
Quality 6.3 (contamination)				
No Disconfirming Information	65			
Quality 6.8 (attrition)				
No disconfirming Information	60			
Quality 7.3 (reporting)				
No disconfirming Information	65			

Subsidized Lunch Status	25			
Total Classrooms	39	6	6.83	9.28
Total Conditions	68	2	2.34	0.61
Total Reading Comprehension Measures	68	1	1.65	1.16
Total Schools	66	4	3.45	6.28
Used Statistical Control or Matching	37			
%White	39	36.00	38.29	27.88
%Black	38	24.97	32.26	26.69
%Asian	26	2.59	3.29	3.34
%Latino	26	23.50	30.03	29.15
%Other	24	0	3.45	6.28

Note. ELL = English language learners; Grade (Min) = the lowest grade level in the sample; Grade (Max) = the highest grade level in the sample; LEP = Limited English Proficiency; N_{cells} = the total amount of relevant cells of data possible; n_{cells} = the amount of relevant data associated with the variables; Quality 6.3 (contamination) = lack of contamination between conditions (no disconfirming information); Quality 6.8 (attrition) = low rates of attrition across groups (no disconfirming information); Quality 7.3 (reporting) = nonselective reporting of significant results (no disconfirming information); Treated Control – SD = Treated Control – Same Domain; Treated Control – DD = Treated Control – Different Domain.

In Table 1, control conditions were typically Business as Usual ([BAU] n_{cells} = 53). A smaller number of conditions described when the control group received interventions in the same domain of reading (i.e., treated control groups – same domain [n_{cells} = 11]). Fewer studies (n_{cells} = 4) were identified as other. Conditions labeled as other provided supplemental materials from the intervention to the control group but lacked instruction (e.g., Boulware-Gooden, Carreker, Thornhill, & Joshi, 2007). The

majority (n_{cells} = 49) of studies were published since 2010. Study quality ratings provided evidence suggesting high quality as it related to lack of contamination between conditions (i.e., n_{cells} = 65 had no disconfirming information), low rates of attrition across groups (i.e., n_{cells} = 60 had no disconfirming information), and nonselective reporting of significant results (i.e., n_{cells} = 65 had no disconfirming information).

Table 2 contains descriptive information for the assessment characteristics. Of note, these variables also had more cells of information than studies. Multiple studies used more than one assessment, leading to a total of N_{cells} = 97. As such, reliability evidence was rarely reported (n_{cells} = 23) in studies. Data regarding assessment administration fidelity was not reported by any included authors.

Table 2

Study-level Assessment Characteristics (N_{cells} = 97)

Variable	n_{cells}	%
Administration Size		
Individual	52	54
Class or Large Group	26	27
Small Group	14	14
Other	5	5
Measure Description	2	2
Cloze	12	12
Informal Reading Inventory	3	3
Maze	4	4
Multiple Choice	19	20
Retell/Summary	1	1

Sentence Verification	56	58
IQ or Achievement		
Response Format		
Multiple Choice	35	36
Cloze	26	27
Maze	4	4
True/False	1	1
Open-Ended Questions	11	11
Summary/Retell	20	21
Picture Selection	0	0
Other	0	0
Score Reliability		
Reported	23	24
Standardized Measure	70	72

Note. N_{cells} = the total amount of relevant cells of data; n_{cells} = the amount of relevant data associated with the variables.

Table 3 contains descriptive information for the intervention characteristics. Of note, these variables also had a greater number of cells than there were studies. This was because one study (i.e., Connor et al., 2018) had two different intervention conditions. Thus, there was a total of N_{cells} = 67 cells of information. As such, few studies (n_{cells} = 25) provided sufficient information to determine the curriculum that all students receive. Treatment fidelity was assessed in n_{cells} = 49 studies, implementers were trained in n_{cells} = 44 studies, and n_{cells} = 32 provided continued consultative supports throughout implementation.

Table 3

Study-level Intervention Characteristics (N_{cells} = 68)

Variable	n_{cells}	*Median*	*M*	*SD*
Curriculum Described	25			
Fidelity of Treatment Reported	49			
Implementer				
Classroom Teacher	33			
Student Teacher	1			
Researcher	20			
Clinician	0			
Special Education Teacher	1			
Parent	0			
Peer	1			
Other	11			
Implementers (Total)	46	5.5	9.67	14.10
Intervention Setting				
Pullout	30			
Classroom	31			
Tutorial	1			
Not Reported	4			
Length of Training (hours)	44	4.5	9.55	10.90
Minutes per Session	56	30	37.68	16.42
Number of Weeks	55	12	14.95	15.35
Sessions per Week	50	4	3.64	1.32
Continued Consultation	32			
Graphic Organizer	19			
Improving Background Knowledge	20			
Inference Instruction	19			
Multicomponent	33			

Other	58
Self-monitoring	18
Signal Words	7
Strategy Instruction	46
Technology	14
Text Structure Instruction	24
Vocabulary Component	35
Writing Component	25

Note. N_{cells} = the total amount of relevant cells of data; n_{cells} = the amount of relevant data associated with the variables.

This section summarized the study-level, assessment, and intervention descriptive characteristics associated with sample of studies included in the meta-analysis. The subsection that follows describes how the sample of studies met assumptions for meta-analysis.

Model Assumptions. Meta-analyses in education should address the extent to which the data met assumptions for statistical analysis (Harwell, 2008). Most commonly, studies address publication bias and outliers. In addition, some meta-analyses also discuss independence. As such, information regarding publication bias, outliers, and independence was provided first, followed by homogeneity of variance and normality.

Publication bias. To begin, publication bias is sampling error that affects the results. Due to the file drawer problem, it's possible that effect sizes are positively biased because studies with nonsignificant or low magnitude findings are less likely to be published (Card, 2012). The current study examined publication bias through the use of visual analysis of funnel plots. The funnel plots used plotted the effect sizes for RQ 1 and

RQ 2 against their standard errors. Asymmetry between the two halves of the plot would suggest publication bias, while a roughly symmetric distribution would support similar representation of studies. Findings and the extent that publication bias was suspected in the current sample are described below.

As described above, funnel plots were used to understand publication bias. Figures 2 and 3 present funnel plots for the posttest (Figure 2) and growth (Figure 3) effect sizes plotted against standard errors.

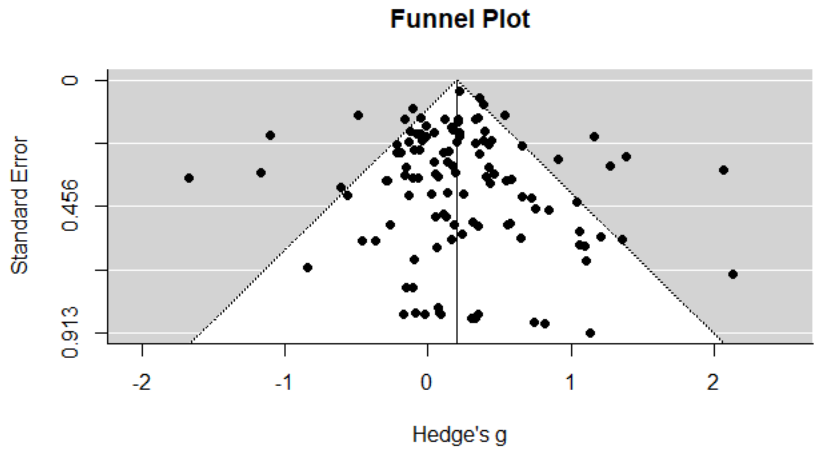

Figure 2. Funnel plot of all posttest effect size estimates compared to their standard errors.

The funnel plot for posttest differences is available in Figure 2. When plotting posttest effect sizes against standard errors, the funnel plot was fairly symmetric with positive and negative effect sizes across varying standard errors. The symmetry on both sides of the funnel plot depicted in Figure 2 suggest similar representation of studies with

small and large standard errors across effect size estimates. As such, given that the funnel plot shows a fairly symmetrical distribution, there was minimal risk of publication bias in the posttest effect sizes.

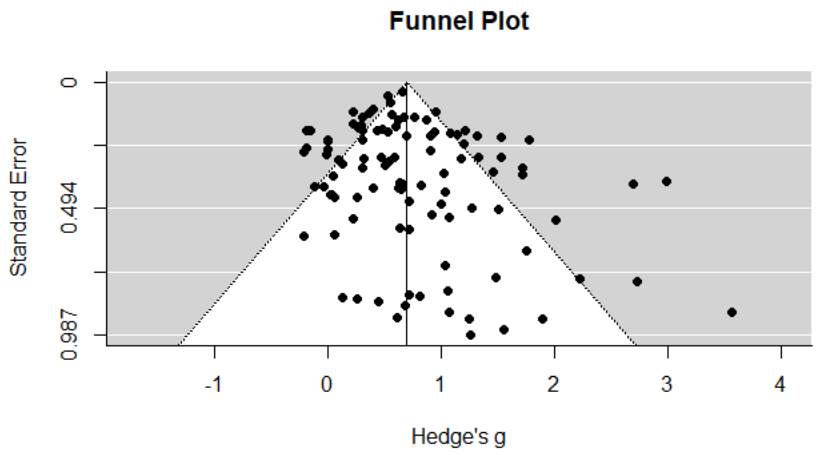

Figure 3. Funnel plot of all within-group growth effect size estimates compared to their standard errors.

The funnel plot for growth effect sizes is represented in Figure 3. The funnel plot appeared less symmetrical, with growth effect sizes plotted against standard errors. Ultimately, despite the asymmetry of the funnel plot, the risk of publication bias was deemed low. The rationale is described below.

The nature of the growth effect size was used to determine that there was minimal risk of publication bias despite some asymmetry in funnel plot depicted in Figure 3. Notably, the mean effect size for the growth effect sizes was closer to one, compared to the mean posttest effect size shown in Figure 2 which was closer to zero. In addition, a

notable area of asymmetry was on the left side of the graph, which represented values that were negative and approaching zero. Thus, conclusions were that the funnel plot for growth effect sizes estimates tended to be above zero, and less likely to be null or negative. The risk of publication bias was decidedly low given that study-specific effect sizes were calculated using only the intervention conditions. It was reasonable to expect these conditions to change over time when measured from pretest to posttest due to factors such as maturation and the effects of the curriculum (e.g., Campbell & Stanley, 1963; Lemons et al., 2014; Swanson et al., 2017). As such, there was a minimal risk of publication bias given that the skew appeared to be a natural artifact of the effect sizes.

Independence of effect sizes. The next assumption explored was independence of effect sizes (Harwell, 2008). Scammacca et al. (2014) examined differing ways to adjust for dependent effect sizes since it is often an issue in intervention research and should be addressed. Two types of dependent effect sizes are represented by studies with (a) multiple intervention conditions and/or (b) multiple measures or dependent variables (Harwell, 2008). In the present study, $m = 2$ effect sizes included in the meta-analysis were from a study with multiple treatment conditions; thus, they were associated with multiple treatments. Likewise, $k = 22$ studies administered more than one measure, and they were associated with multiple measures or dependent variables. In addition, $m = 58$ effect sizes originated from studies that reported results based on multiple populations (e.g., different grades or stratifications of students).

The present study adjusted the sampling variance to account for dependent effect sizes using Equation 1:

$$w_{adjusted} = \frac{1}{\Delta_j * m_j} \quad (1),$$

where $w_{adjusted}$ was the adjusted weight for the effect size, and it was calculated by taking the inverse of the sampling variance (Δ_j) multiplied by the total number of effect sizes (m) within that particular study (j). Adjusting the sampling variance prevented studies from being counted multiple times within the meta-analysis. The sampling variance was multiplied because it adds more dispersion, which overall decreases the weight and precision of a particular effect. Thus, effect sizes originating from the same study shared the weight instead of gaining more. As such, the study also addressed assumptions relevant to independence of effect sizes. The next assumption discussed, as recommended (Harwell, 2008), was normality. Of note, all equations referenced in the Results section are provided in Appendix C.

Two additional factors relevant to assumptions are homogeneity of variance and normality. The present study used skew, kurtosis, and smoothed density plots of effect sizes from both the posttest and growth meta-analyses to examine normality (Harwell, 2008). In the case of the posttest meta-analysis effect sizes, skew was valued at 0.34, kurtosis at 1.91, and visual analysis of density plots suggested the data were approximately normal. In the case of growth effect sizes, skew was valued at 1.28, kurtosis at 2.08, and visual analysis of density plots suggested a slight positive skew. Outliers were examined using boxplots that identified values 1.5 times above or below the upper and lower quartiles, respectively. A total of seven effect sizes met the boxplot criteria for outliers in the posttest sample, and three effect sizes met the boxplot criteria for outliers in the sample in the growth sample. No effect sizes were removed or

winsorized, in order to maintain sample size and because the sample of studies represented reading comprehension interventions. Any anomalies from the norm could be accounted for through the regression analyses.

Homogeneity of variance was evaluated using Levene's test statistic (Howell, 2013). Results suggested that the data adequately met the assumptions for homogeneity of variance for means between groups ($F_{\text{Levene,}} < .001$, $p = .99$) and between pretest and posttest scores ($F_{\text{Levene}} < .001$, $p = .98$).

Finally, power calculations were conducted to determine the likelihood of detecting a significant finding when there truly was an effect (Howell, 2013). Retrospective power analyses were conducted using the sample size, level of heterogeneity, alpha level of .05, and examining for effect sizes of 0.20 returned a power value greater than .99 for both posttest and growth analyses (Harrer et al., 2019).

Altogether, these data were examined to understand the extent to which statistical assumptions were sufficiently met. Evidence suggested data sufficiently met assumptions for homogeneity of error variances. Visual analysis of funnel plots suggested risk of publication bias was low. Data relevant to independence of the effect sizes were reported, and adjustments to error variances were used. Finally, the study was over powered to detect an effect of 0.20 for RQ 1 and RQ 2.

Research Question 1: Posttest Meta-analyses

The goal of RQ 1 was to investigate the overall treatment effect of reading comprehension interventions compared to control conditions. Findings are reported using

Hedge's *g*, which is standardized mean difference with a correction for small sample sizes.

In order to address RQ 1, a meta-analysis was conducted using all available treatment to control group comparisons at posttest effect sizes. These variables were entered in the model without the use of any predictor variables. As such, the first model is an intercept-only meta-analysis; thus, it did not use any predictors. This may also be referred to as a null model. The intercept-only meta-analysis was represented in Equation 2 as,

$$g_j = \gamma_0 + u_j + e_j, \quad (2)$$

where, g_j was the weighted treatment effect of reading comprehension interventions using the formula for weight provided in Equation 1, γ_0 was the intercept, and u_j and e_j were error terms associated with between and within study variance, respectively.

To begin, an overall weighted mean effect size was calculated across all eligible posttest effect sizes. Therefore, the sample included both quasi and experimental designs. Table 4 contains the weighted effect size estimate across all eligible studies. The findings for this particular analysis are under Posttest Analyses, and in the row labeled All Study Designs. The overall treatment effect of reading comprehension interventions on reading comprehension outcomes was statistically significant ($p < .001$). The overall weighted Hedge's *g* effect size was 0.20 (95% CI [0.10, 0.29]), which was a considered small effect using Cohen's (1988) standards.

Table 4

Sensitivity Analyses for Posttest and Growth Meta-analyses

Design	*k*	*m*	*g*	*SE*	*p*	*95% CI*	τ^2	*Q*	I^2
Posttest Analyses									
All Study Designs	66	116	0.20	0.05	<.001	[0.10, 0.29]	.15	< .001	76.19%
Experimental Only	48	93	0.19	0.04	<.001	[0.10, 0.28]	.07	< .001	60.75%
Growth Analyses									
All Study Designs	58	104	0.71	0.06	<.001	[0.59, 0.83]	.26	< .001	86.31%
Experimental Only	41	82	0.69	0.07	<.001	[0.55, 0.82]	.22	< .001	85.20%

Note. All Study Designs describes the inclusion of all eligible study and effect sizes regardless of experimental design used, while Experimental Only includes only eligible study and effect sizes that used experimental designs and did not include studies with quasi-experimental designs. *k* = number of studies; *m* = number of effect sizes; g = Hedge's *g* mean effect size; τ^2 = the estimated between study variance.

Sensitivity analyses. A sensitivity analysis was used to explore if there were differences between intercept-only meta-analytic results when using the full sample of studies (*k* = 66; *m* = 116) compared to intercept-only meta-analytic results using a subsample of studies that only included experimental designs (*k* = 48; *m* = 93). No differences in findings would support the use of effect sizes from both experimental and quasi-experimental designs. Notable differences would support the use of the subsample of studies and effect sizes, including only experimental designs in the remaining meta-analyses of posttest effect sizes.

The meta-analysis using only experimental designs was run similar to the meta-analysis using all quasi and experimental design effect sizes. Thus, the model was still represented by Equation 2 (above) where g_j was the weighted treatment effect of reading

comprehension interventions from the subset of experimental studies, γ_0 was the intercept, and u_j and e_j were error terms associated with between and within study variance, respectively. The results of this model are presented in Table 4, under Posttest Analyses, and in the row labeled, Experimental Only.

The intercept-only meta-analysis using only studies with experimental designs resulted in a statistically significant ($p < .001$) weighted mean effect size that was small in magnitude ($g = 0.19$, 95% CI [0.10, 0.28]). Estimates between the sample using all posttest effect sizes ($m = 116$) and the sample using only experimental designs ($m = 93$) were similar in magnitude.

The 95% confidence intervals for the two weighted mean effect sizes overlapped, which failed to suggest significant differences between the two intervals. Although this approach is susceptible to type I error, this approach failed to suggest that there was a significant difference between the two posttest analysis effect size estimates reported in Table 4. In addition to the weighted mean effect size estimates in Table 4, the between-study heterogeneity statistics are also reported: τ^2, Q, and I^2. Similar to the weighted mean effect size estimates (i.e., Hedge's g), confidence intervals overlapped for the meta-analysis using all available (i.e., from both quasi and experimental designs) posttest effect sizes' τ^2 (95% CI [0.09, 0.22]) and the experiment-only meta-analysis' τ^2 (95% CI [0.03, 0.12]). Furthermore, Q statistics revealed significant levels of heterogeneity in both samples. The sample with both quasi and experimental designs had a heterogeneity value of $I^2 = 76.19\%$, where $I^2 > 75\%$ is considered large (Card, 2012). In contrast, the experiment-only designs had an $I^2 = 60.75$, where $50\% < I^2 < 75\%$ is considered a

medium level of heterogeneity (Card, 2012). Altogether, these results from sensitivity analysis did not indicate differences between findings from the two meta-analyses. Thus, the full sample of studies, which included both quasi and experimental designs, was used to address the remainder of RQ 1.

As indicated by the results describe above, there was enough of variability in the study outcomes (i.e., heterogeneity in effect sizes) to merit the use of meta-regressions to see what explained variation in study outcomes. That is, between-study heterogeneity in the full model of posttest effects was statistically significant with a Q statistic ($p < .001$) and an I^2 value of 76.19%. Those results indicate there is additional variance to explain with models that account for study-specific and intervention-specific characteristics.

Model building using blocks: Defining blocks. As an initial step to explore the variability in study results, relevant characteristics for each study were identified and coded. They related to study design, reader characteristics, intervention characteristics, and assessment characteristics. For the remainder of the paper, the term "blocks" is used to describe these broad categories that were coded and used to define characteristics. These characteristics were entered into meta-regressions at once (i.e., in one block) to identify characteristics that explained differences, or variance, in outcomes across studies. There were two exceptions to all variables in a block being entered at once, but they are described in the next section, Model building using blocks.

Study characteristics: Study design block. The following variables were included in the study design block: control condition, level of randomization, matching, number of

conditions, number of measures, publication type, publication year, and quasi-experimental design. These variables are included in Table 5.

Table 5

Study Design Meta-regressions

	Post (*m* = 116)			Growth (*m* = 86)		
Variable	γ	*SE*	*p*	γ	*SE*	*p*
Control Condition:						
Treated Control	-0.30	0.15	.041	0.15	0.19	.448
Other	0.30	0.27	.256	-0.14	0.37	.701
Level of Randomization:						
Student	0.20	0.33	.556	-0.69	0.43	.107
Classroom	0.30	0.32	.359	-0.56	0.40	.164
School	0.47	0.37	.205	-0.04	0.48	.931
Matching	-0.03	0.12	.806	-0.18	0.16	.278
Number of Conditions	-0.05	0.09	.593	0.25	0.16	.115
Number of Measures	**0.17**	**0.05**	**<.001**	0.05	0.05	.347
Publication Type:						
Dissertation	0.07	0.13	.604	-0.03	0.19	.866
Other	0.06	0.19	.751	-0.10	0.26	.705
Publication Year	-0.02	0.01	.185	0.02	0.02	.200
Quasi-Experiment	0.26	0.32	.423	-0.29	0.41	.474

Note. Variables in bold remained significant in reduced block analysis and were included in the final model. *m* = number of effect sizes; γ = the regression coefficient associated with variable.

Study characteristics: Intervention characteristics block. The following variables were included in the model under the intervention characteristics block: continued consultation; implementer; treatment fidelity; intervention setting; and intervention components of graphic organizers, improving background knowledge, inference instruction, multicomponent, other, self-monitoring, signal words, strategy instruction, technology, text structure instruction, vocabulary components, and writing components. These variables are included in Table 6.

Table 6

Intervention Meta-regressions

	Post (*m* = 116)			Growth (*m* = 86)		
Variable	γ	*SE*	*p*	γ	*SE*	*p*
Intervention Context						
Continued Consultation	**-0.36**	**0.10**	**<.001**	0.18	0.14	.202
Implementer:						
Researcher	-0.01	0.17	.95	0.17	0.24	.479
Classroom Teacher	0.15	0.15	.33	0.34	0.19	.082
Other	0.03	0.14	.83	0.11	0.18	.533
Treatment Fidelity	**0.35**	**0.12**	**.003**	-0.15	0.17	.353
Intervention Setting:						
Classroom	NA	NA	NA	**1.44**	**0.51**	**.005**
Pullout	NA	NA	NA	**1.49**	**0.50**	**.003**
Tutorial	NA	NA	NA	1.30	0.79	.100
Intervention Type						
Graphic Organizer	**0.61**	**0.13**	**<.001**	**0.45**	**0.18**	**.011**

Improving Background Knowledge	-0.29	0.11	.011	-0.23	0.16	.170
Inference Instruction	0.21	0.12	.080	0.32	0.18	.074
Multicomponent	0.16	0.12	.182	0.18	0.16	.252
Other	-0.31	0.17	.059	-0.47	0.26	.065
Self-Monitoring	-0.21	0.12	.075	0.20	0.16	.225
Signal Words	0.16	0.21	.447	-0.06	0.29	.844
Strategy Instruction	0.15	0.11	.188	0.15	0.16	.323
Technology	**-0.40**	**0.12**	**.001**	0.01	0.18	.948
Text Structure Instruction	0.15	0.12	.233	0.30	0.18	.088
Vocabulary Component	-0.14	0.12	.241	0.13	0.18	.447
Writing Component	-0.26	0.14	.059	-0.20	0.19	.305

Note. Variables in bold remained significant in reduced block analysis and were included in the final model. *m* = number of effect sizes; γ = the regression coefficient associated with variable.

Study characteristics: Assessment characteristics block. The following variables were included in the assessment characteristics block: administration size; administrator; standardized measure; and the response formats of cloze, maze, multiple choice, open-ended questions, summary or retell, and true/false. These variables are included in Table 7.

Table 7

Assessment Meta-regressions

	Post (*m* = 116)			Growth (*m* = 86)		
Variable	γ	*SE*	*p*	γ	*SE*	*p*
Intercept	0.42	0.11	<.001	0.80	0.14	<.001
Administration Size	-0.06	0.05	.269	-0.10	0.07	.194
Administrator						
Computer	-0.43	0.24	.073	-0.12	0.43	.782
Other	0.10	0.37	.791	-0.17	0.44	.707
Teacher	0.07	0.19	.731	-0.22	0.30	.460
Standardized Measure	-0.22	0.12	.059	<0.00	0.16	.977
Response Format						
Univariate Regressions:						
Intercept	0.24	0.06	<.001	0.72	0.08	<.001
Cloze	-0.14	0.11	.179	-0.02	0.14	.854
Intercept	0.20	0.05	<.001	0.72	0.07	<.001
Maze	-0.002	0.28	.994	-0.20	0.33	.540
Intercept	0.25	0.06	<.001	0.78	0.08	<.001
Multiple Choice	-0.13	0.10	.193	-0.18	0.13	.175
Intercept	0.20	0.05	<.001	0.71	0.07	<.001
Open-Ended Questions	-0.04	0.16	.794	0.05	0.22	.817
Intercept	0.14	0.05	.004	0.66	0.07	<.001
Retell/Summary	**0.44**	**0.14**	**.002**	**0.43**	**0.19**	**.026**
Intercept	0.20	0.05	<.001	0.71	0.06	<.001
True/False	0.22	0.44	.624	0.33	0.49	.496

Note. Variables in bold remained significant in reduced block analysis and were included in the final model. Intercepts are included for univariate regressions to provide further clarity of the overall effect of the response format in relation to other formats. *m* = number of effect sizes; γ = the regression coefficient associated with variable.

Study characteristics: Reader characteristics block. The following variables were included in the reader characteristics block: lowest grade level in the sample (i.e., lowest grade [minimum]) and highest grade level in the sample (i.e., grade [maximum]). These variables are included in Table 8.

Table 8

Reader Characteristic Meta-regressions

	Post (*m* = 116)			Growth (*m* = 86)		
Variable	γ	*SE*	*p*	γ	*SE*	*p*
Grade (Minimum)	0.07	0.05	.208	0.05	0.08	.501
Grade (Maximum)	-0.11	0.06	.054	-0.14	0.09	.110

Note. *m* = number of effect sizes; γ = the regression coefficient associated with variable; Grade (Minimum) = lowest grade level in the sample; Grade (Maximum) = highest grade level in the sample.

Study characteristics: Study quality block. The following variables were included in the study quality block: quality 6.3 to represent studies with varying evidence of contamination between treatment and control conditions, quality 6.8 to represent low attrition (defined by <30% attrition in a 1-year study), and quality 7.3 to represent non-

selective reporting of treatment outcomes across measures. These variables are included in Table 9.

Table 9

Study Quality Meta-regressions

	Post (*m* = 116)			Growth (*m* = 86)		
Variable	γ	*SE*	*p*	γ	*SE*	*p*
Quality 6.3	**0.95**	**0.36**	**.008**	0.43	0.74	.561
Quality 6.8	0.27	0.18	.139	-0.32	0.24	.181
Quality 7.3	<0.01	0.29	.996	<0.01	0.52	.994

Note. Variables in bold remained significant in reduced block analysis and were included in the final model. *m* = number of effect sizes; γ = the regression coefficient associated with variable; Quality 6.3 = lack of contamination between conditions (no disconfirming information); Quality 6.8 = low rates of attrition across groups (no disconfirming information); Quality 7.3 = nonselective reporting of significant results (no disconfirming information).

Model building using blocks: Approach. Each of the blocks were used to explain differences between studies. This process was done using meta-regression. Meta-regression is the use of regression methods for a meta-analysis. Regression is the use of statistical procedures to explain the association between a dependent variable and moderator variable(s). In the meta-regression, the dependent variable is the effect size, and the moderator variables are at the study level (Borenstein et al., 2009). Specifically, the present study used weighted least squares regression with maximum likelihood

estimation to assign values to study-level variables that may be associated with significant differences in posttest effect sizes (Raudenbush & Bryk, 2002).

As such, two meta-regressions were run for each block. In each meta-regression, relevant predictors from the block were included. Each of the regression models adhered to the following core structure represented in Equation 3:

$$g_j = \gamma_0 + \gamma_1 W_{1j} + \gamma_2 W_{2j} + \ldots + \gamma_s W_{sj} + u_j + e_j \; . \qquad (3),$$

where, γ_0 is the intercept, $\gamma_{1 \ldots} \gamma_s$ represents the value of the model estimated coefficient, and W_{sj} represents the specific study level variable. Finally, u_j and e_j represent error terms associated with between and within study variance, respectively.

Coded variables were entered as predictors in meta-regressions according to their block. The procedure for the meta-regression analyses by block was that variables were entered collectively – as opposed to incrementally – into the meta-regression. Next, variables that were significant were identified and then entered in a reduced model of variables for that particular block. Any variable that maintained significance was used to construct a final model. Thus, the purpose of the block meta-regression analyses was to purposively select variables from each block that explained significant levels of heterogeneity in effect sizes. It was expected that response formats from the assessment block would be included in the final model and maintain significance when controlling for other block characteristics.

Model building using blocks: Results for posttest. Tables 5, 6, 7, 8, and 9 contain information regarding findings for the study design, intervention characteristics, assessment characteristics, reader characteristics, and study quality blocks respectively.

In each table, the block analysis for the posttest effect sizes are provided under the term, Post, which is associated with the meta-analysis of posttest effect sizes associated with RQ 1. The term, Growth, was used to represent the block analyses for the growth effect sizes used in RQ 2. The model estimated coefficient for each variable (γ_s), its standard error, and the significance value for each coefficient are provided.

The table also contains the variables that were significant after being entered in a reduced block analysis, which was a meta-regression of block variables that were significant when all other variables within the block were entered in the first meta-regression. Variables that maintained significance were then entered into the full model. In each table, those variables are in bold. In summary, two rounds of block analyses were conducted to identify variables that explained significant variation in effect sizes. The first round of block analysis was used to identify significant predictors among relevant block variables. The second analysis identified which variables maintained significance in a reduced model. This was done for each of the blocks: study design, intervention characteristics, assessment characteristics, reader characteristics, and study quality. The results are described below.

Block analysis: Study design. Beginning with study design, all relevant variables were entered into the model. The model included the following variables shown in equation 4:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Control Condition: Treated Control } j} + \gamma_2 W_{\text{Control Condition: Other } j} + \gamma_3 W_{\text{Level of randomization: Student } j} + \gamma_4 W_{\text{Level of randomization: Classroom } j} + \gamma_5 W_{\text{Level of randomization: School } j} +$$

$$\gamma_6 W_{\text{Matching } j} + \gamma_7 W_{\text{Number of Measures } j} + \gamma_8 W_{\text{Publication Type: Dissertation } j} + \gamma_9 W_{\text{Publication Type: Other } j} + \gamma_{10} W_{\text{Publication Year } j} + \gamma_{11} W_{\text{Quasi-experiment } j} + u_j + e_j \quad (4).$$

The results of the model, specifically the γ_s, are available in Table 5. Of the variables included in the model, only two were significant in the block test. Those variables were number of (reading comprehension) measures used and treated control conditions (compared to BAU). Nonsignificant variables were removed, resulting in the following reduced model shown in Equation 5:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Control Condition: Treated Control } j} + \gamma_2 W_{\text{Number of Measures } j} + u_j + e_j \quad (5).$$

In the reduced model, including only design variables that were significant, number measures maintained significance at the .05 level, while treated control condition did not. Therefore, number of reading comprehension measures would be included in the final model while number of measures was not. There was a medium amount of unexplained heterogeneity in this model, Q (112) = 324.48, $p < .001$, $I^2 = 71.21\%$.

Block analysis: Intervention characteristics. The intervention block was initially separated into two regression analyses given the number of variables and the focus of the study. Characteristics of the intervention's context were included in their own analysis, and intervention types were included in a separate analyses.

These analyses are represented in Table 6. The first analysis used variables from the intervention characteristics block related the intervention context. They are represented in Equation 6:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Continued Consultation } j} + \gamma_2 W_{\text{Implementer: Researcher } j} + \gamma_3 W_{\text{Implementer: Classroom Teacher } j} + \gamma_4 W_{\text{Implementer: Other } j} + \gamma_4 W_{\text{Treatment Fidelity } j} + \gamma_5 W_{\text{Intervention Setting: Classroom } j} + \gamma_6 W_{\text{Intervention Setting: Pullout } j} + \gamma_7 W_{\text{Intervention Setting: Tutorial } j} + u_j + e_j \quad (6).$$

In relation to intervention context, whether interventionists received continued consultation during implementation and treatment fidelity were the only significant variables.

The second analysis used the following structure shown in Equation 7:

$$g_j = \gamma_0 + \gamma_1 W_{\text{ Graphic Organizer } j} + \gamma_2 W_{\text{ Improving Background Knowledge } j} + \gamma_3 W_{\text{ Inference Instruction } j} + \gamma_4 W_{\text{ Multicomponent } j} + \gamma_5 W_{\text{ Other } j} + \gamma_6 W_{\text{ Self-monitoring } j} + \gamma_7 W_{\text{ Signal Words } j} + \gamma_8 W_{\text{ Strategy Instruction } j} + \gamma_9 W_{\text{ Technology } j} + \gamma_{10} W_{\text{ Text Structure Instruction } j} + \gamma_{11} W_{\text{ Vocabulary Component } j} + \gamma_{11} W_{\text{ Writing Component } j} + u_j + e_j \quad (7).$$

Interventions that included graphic organizers, improved background knowledge, and technology were significant, when accounting for other types of interventions. As can be observed in Table 6, several of the aforementioned variables were negatively associated with differences at posttest. Small, negative relationships were observed for continued consultation (γ = -0.36) and technology (γ = -0.40). These negative weighted effect size estimates suggest lower performance in the treatment group compared to the control, on average, and controlling for other intervention variables.

Significant intervention variables were then combined in a reduced model of the block. The model was represented in Equation 8 as,

$$g_j = \gamma_0 + \gamma_1 W_{\text{ Continued Consultation } j} + \gamma_2 W_{\text{ Treatment } \textit{Fidelity } j} + \gamma_3 W_{\text{ Graphic Organizer } j} + \gamma_4 W_{\text{ Improving Background Knowledge } j} + \gamma_5 W_{\text{ Technology }} + u_j + e_j \quad (8).$$

The model resulted in the following variables being significant predictors of group differences at posttest: continued consultation, treatment fidelity, graphic organizers, and technology. The significant variables are represented in bold in Table 6. The improving background knowledge variable was excluded from being included in the final model, as it was nonsignificant, γ = -0.14 (*SE* = 0.10), p = .17. Altogether, there was a medium amount of unexplained heterogeneity in the model, Q (110) = 271.99, p < .001, I^2 = 66.63%.

Assessment characteristics. Similar to the intervention block, the assessment block was divided into two areas: response format and assessment context characteristics. The results from both of these analyses are reported in Table 7. The model for assessment context characteristics is shown in Equation 9:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Administration Size } j} + \gamma_2 W_{\text{Administrator } j} + \gamma_3 W_{\text{Standardized Measure } j} + u_j + e_j \quad (9).$$

None of the assessment context characteristic variables were significant in the model. In relation to response formats, each format was dummy coded and entered as univariate regressions. For example, the model for cloze is shown in Equation 10 as,

$$g_j = \gamma_0 + \gamma_1 W_{\text{Cloze } j} + u_j + e_j \quad (10).$$

Of the response formats, retell/summary format was the only significant variable.

A block analysis for assessment was conducted including the two significant variables. It is shown in Equation 11 as,

$$g_j = \gamma_0 + \gamma_1 W_{\text{Summary } or \text{ } Retell \text{ } j} + u_j + e_j \quad (11).$$

In the block model with only previously significant variables included, the retell/summary format maintained significance. The assessment block resulted in a medium level of heterogeneity, Q (114) = 312.24, p = .002, I^2 = 72.66%.

Reader characteristics. The reader block consisted of only two variables that had sufficient data. As such, the model is represented in Equation 12 as,

$$g_j = \gamma_0 + \gamma_1 W_{\text{Grade (Minimum) } j} + \gamma_2 W_{\text{Grade (Maximum) } j} + u_j + e_j \qquad (12).$$

These results are available in Table 8. In the reader characteristic block, minimum and maximum grade level were not significant. As presented in Table 1, variables related to race, free and reduced lunch status, number of schools, number of classrooms, special education status, and English language learning status were reported at too low of a rate to provide meaningful information. They were excluded from the block analysis as a result. Given that neither minimum and maximum grade were the only values consistently reported across studies and that neither was significant, no additional analyses for reader characteristics were conducted within the block. Thus, neither variable was included in the final regression model incorporating significant variables from each block.

Study quality. Finally, the study quality block is displayed in Table 9. The model is also represented in Equation 13 as,

$$g_j = \gamma_0 + \gamma_1 W_{\text{ Quality 6.3 } j} + \gamma_2 W_{\text{ Quality 6.8 } j} + \gamma_3 W_{\text{ Quality 7.3 } j} + u_j + e_j \qquad (13).$$

One variable was significant, and it pertained to treatment contamination (i.e., quality indicator 6.3). The reduced block model, which included the study quality variable related to treatment contamination (6.3) was associated with a medium level of the

heterogeneity between weighted mean effect sizes, Q (114) = 306.29, $p < .001$, I^2 = 73.31%.

Final Step: Full Model. The significant variables from each reduced block analysis were used to construct a final model. These variables are represented in bold in Tables 5, 6, 7, 8, and 9. In addition, significant variables from each block are also represented in Table 10. The following variables were significant in the full model: continued consultation, graphic organizers, self-monitoring, retell/summary, and quality 6.8 (measured treatment contamination). The variables may also be expressed in terms of their meta-regression in equation 14,

$$g_j = \gamma_0 + \gamma_1 W_{\text{Continued Consultation } j} + \gamma_2 W_{\text{Graphic Organizer } j} + \gamma_3 W_{\text{N Measures } j} + \gamma_4 W_{\text{Retell/summary } j} + \gamma_5 W_{\text{Technology } j} + \gamma_6 W_{\text{Quality 6.3 } j} + u_j + e_j \quad (14).$$

There were positive and negative estimates associated with the variables. Continued consultation (γ_1 = -0.19) and technology (γ = -0.18) were associated with negative effects. Interventions that incorporated graphic organizers (γ = 0.29) and response formats that used retell/summary (γ = 0.39) were associated with small, positive weighted mean effect size differences between treatment and control at posttest. Quality indicator 6.3 (Contamination; γ = 1.07) was associated with large, positive weighted mean effect size differences between treatment and control at posttest. Nonsignificant variables were the number of reading comprehension measures, the use of technology in the intervention, and treatment fidelity. Overall, the final model was associated with a statistically significant and medium amount of heterogeneity, Q (108) = 213.75, $p < .001$, I^2 = 55.25%.

Table 10

Full Model Meta-regression Using all Relevant Block Variables

	Post [a] (*m* = 116)			Growth [b] (*m* = 86)		
Variable	γ	*SE*	*p*	γ	*SE*	*p*
Intercept	-2.13	0.64	<.001	-0.39	0.38	.301
Continued Consultation	-0.19	0.09	.039			
Graphic Organizer	0.29	0.10	.005	0.26	0.14	.073
Intervention Setting:						
Pullout				0.98	0.38	.009
Classroom				0.97	0.37	.010
Number of Measures	0.04	0.04	.240			
Retell/Summary	0.39	0.14	.004	0.38	0.19	.046
Technology	-0.18	0.10	.063			
Treatment Fidelity	0.19	0.11	.068			
Quality 6.3	1.07	0.32	<.001			
(Contamination)						

Note. *m* = number of effect sizes; γ = the regression coefficient associated with variable; τ^2 = the estimated between study variance; Quality 6.3 (Contamination) = lack of contamination between conditions (no disconfirming information).
[a] $\tau^2 = 0.08$ (*SE* = 0.02), *Q* (108) = 213.75, $p < .001$, $I^2 = 55.25\%$. [b] $\tau^2 = 0.24$ (*SE* = 0.05), *Q* (96) = 357.92, $p < .001$, $I^2 = 82.05\%$

In summary, these findings suggest that continued consultation and interventions that used technology were significant, negative predictors of the posttest differences between treatment and control conditions. Likewise, the quality indicator associated with treatment contamination as well as graphic organizers and retell/summary response formats were statistically significant positive predictors of posttest differences, where

intervention conditions receiving intervention with these components scored tended to scored at least 0.25 of a standard deviation higher than the control group at posttest, on average. The following variables were no longer significant after accounting for the previously mentioned variables: number of reading comprehension measures, treatment fidelity, graphic organizers, and technology. In addition, after accounting for these variables, there was a medium amount of heterogeneity present (I^2 = 55.25%).

Research Question 2

RQ 2 asked, what is the observed aggregate treatment effect for all reading comprehension interventions delivered to students in elementary school from pretest to posttest? The procedures for analysis and model building were identical to those used for the first research question. However, growth effect sizes from pretest to posttest were used instead of differences between treatment and control at posttest. Each model followed the same core structure as shown in Equation 4. Findings are still reported using Hedge's *g*.

In order to address RQ 2, a meta-analysis was conducted that used all available effect sizes, which included those from both quasi and experimental designs. The variables were entered into an intercept-only model, which is represented in Equation 2. The intercept-only model returned an overall weighted mean effect size that was calculated across all eligible growth effect sizes from quasi and experimental designs. The results of the meta-analysis for RQ 2 are presented in Table 4 under Growth Analyses and in the row labeled, All Study Designs. The estimated weighted mean Hedge's *g* effect size for growth from pretest to posttest across quasi and experimental

designs was significant. The weighted mean effect size was 0.71 (95% CI [0.59, 0.83]), which is considered a moderate effect using Cohen's (1988) standards.

Sensitivity analyses. Similar to RQ 1, a sensitivity analysis was conducted for RQ 2 as well. The sensitivity was used to examine differences between the intercept-only meta-analysis that used both quasi and experimental designs (k = 58; m = 104) and the intercept-only meta-analysis that used only experimental designs (k = 41; m = 82). In short, the sensitivity analysis was conducted to determine if the exclusion of quasi-experimental designs was associated with notable differences in meta-analytic outcomes. It was suspected that major differences between the two models suggest that the quasi-experimental studies biased the overall weighted mean effect size estimate.

For the purposes of sensitivity analysis, an intercept-only meta-analysis was conducted using only effect sizes from experimental designs. The model was constructed similarly to the meta-analysis using all quasi and experimental design effect sizes. Therefore, the model is also represented by Equation 2. The results of the analysis are provided in Table 4, under Growth Analyses, and in the row labeled, Experimental Only. Findings suggested that the intercept-only meta-analysis using only effect sizes from experimental designs resulted in a weighted mean effect size estimate that was also statistically significant ($p < .001$) and moderate in size (γ_0 = 0.69, 95% CI [0.55, 0.82]).

The confidence intervals for the weighted mean effect sizes produced in each of the two meta-analyses were compared. Confidence intervals for the weighted mean effect sizes overlapped for the full model with quasi and experimental designs (95% CI [0.59, 0.83]) and the subset model with only experimental designs (g = 0.69, 95% CI [0.55,

0.82]). Likewise, the confidence intervals associated with the τ^2 estimates between the overall (95% CI [0.12, 0.33]) and experimental design-only (95% CI [.15, .36]) models overlapped. As previously stated, this approach is susceptible to alpha error; however, the approach failed to suggest there was a significant difference between the two growth intercept-only weighted mean effect size and τ^2 estimates.

Altogether, the results of the sensitivity analysis failed to suggest differences between findings from the two meta-analyses. In short, the two intercept-only models appeared similar. The weighted mean effect size estimates (γ_0), statistical significance, and level of heterogeneity (I^2) remained comparable in the model using only experimental designs. As a result, the sample of studies included in the meta-analysis was not reduced to only experimental designs. The full sample of studies, which included both quasi and experimental designs, was used to explore the remainder of RQ 2.

The first step was to determine whether the full sample of growth effect sizes was reasonable to use to address RQ 2. Sensitivity analysis failed to discourage this approach, and the full sample of studies was included. The next step was to evaluate whether there was a significant level of heterogeneity worth exploring through meta-regression. As previously stated, *Q* statistics revealed a significant heterogeneity in the sample of effect sizes, and that amount of heterogeneity was valued at $I^2 = 86.31$, which is considered a large level of heterogeneity (Card, 2012). Therefore, next steps involved using study-level variables that could account for the large between study differences in effect sizes.

The procedures for categorizing and selecting study-level variables followed the same block analysis procedures from RQ 1. A review of blocks (i.e., definition and

contents) is available in the sections, Model building using blocks: Defining blocks and Model building using blocks: Approach. However, blocks is a term used to describe the broad categories of related study level coded characteristics. There were five blocks: the study design block, intervention characteristics block, assessment characteristics block, reader characteristics block, and study quality block. Significant variables were identified in each block trough meta-regressions, which followed the general formula shown in Equation 3. Each variable's model estimated coefficient, standard error, and its *p*-values are displayed under the columns labeled Growth in Tables 5, 6, 7, 8, and 9 for the study design, reader characteristic, intervention, assessment, and study quality blocks, respectively.

Variables that were significant in the initial meta-regression by block were entered in a reduced model. For example, the study quality block has three predictors. If only two were significant in the first meta-regression including all of them, only the two that were significant would be included in the next meta-regression. The variables that maintained significance in the reduced meta-regression were included in the final model. The rationale for this approach was that it would allow for a final meta-regression model that included purposively selected statistically significant variables from a variety of blocks. The variables would also serve as controls for variables related to response format. The block analyses for the growth meta-analysis are reported in the section below.

Model building using blocks: Results for growth. No variables were significant in the block test of study design characteristics (Equation 4; Table 5). Similarly, no

variables were significant in the block test of reader characteristics (Equation 12; Table 8). Furthermore, variables within study quality were not associated with significant effects (Equation 13; Table 9). As such, no reduced models were created, and variables from the study design, reader characteristics, or study quality blocks were not included in the final regression model.

Block analysis: Intervention characteristics. In relation to the intervention block analysis for growth, the variables were also divided as in RQ 1. These analyses are represented in Table 6, with the first set of analyses being done on the variables beneath the underlined label, Intervention Context. The model for intervention context variables was represented in Equation 15:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Continued Consultation } j} + \gamma_2 W_{\text{Implementer } j} + \gamma_3 W_{\text{Treatment Fidelity } j} + \gamma_5 W_{\text{Intervention Setting: Classroom } j} + \gamma_6 W_{\text{Intervention Setting: Pullout } j} + \gamma_7 W_{\text{Intervention Setting: Tutorial } j} + u_j + e_j \quad (15).$$

It is also represented in Table 6. The following variables returned significant values in the first analysis: classroom as the intervention setting and pullout as the intervention setting. Both variables had a positive relationship with treatment growth over time.

The remaining variables in the intervention characteristics block were then run. Those variables are represented in table 6 under the label, Intervention Type. They were also represented in Equation 7. In relation to intervention characteristics, graphic organizers was the only significant predictor.

A reduced model for intervention characteristics was constructed, including all variables that were previously significant (i.e., intervention setting – classroom,

intervention setting – pullout, and graphic organizer). The model is represented in Equation 16:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Intervention Setting: Classroom } j} + \gamma_2 W_{\text{Intervention Setting: Pullout } j} + \gamma_3 W_{\text{Graphic Organizer } j} + u_j + e_j \quad (16).$$

In the reduced model, each variable remained significant. This was indicated in table 6, where significant variables were bolded. As a result, intervention setting – classroom, intervention setting – pullout, and graphic organizers were included in the final model. In addition, the model had a large amount of heterogeneity, Q (97) = 374.22, $p < .001$, I^2 = 82.63%.

Assessment block. In relation to the assessment block, several meta-regressions were conducted. The first represented assessment context characteristics. The model for the growth effect sizes was written as shown in Equation 9. Each of the variables in the model, administration size, administrator, and standardization of the measure were nonsignificant within the block (Table 7). In relation to response formats, univariate regression analyses were conducted. For example, Equation 11 represents the model for retell/summary. Similar to RQ 1, retell/summary was the only response format that was identified as significant. The reduced block analysis containing significant variables only included the retell/summary predictor, which is also represented by Equation 11. The retell/summary format was statistically significant (γ = 0.50, $p <$ .01). A large level of heterogeneity was present in the model, Q (99) = 385.35, $p < .001$, I^2 = 84.38%.

Full model for growth meta-analysis. The significant variables from the block analyses were used for the final regression model for pretest to posttest growth within the

treatment group. The model is represented in Table 10 in the series of columns under Growth. The model included the following variables: graphic organizer, intervention setting – pullout, intervention setting – classroom, and retell/summary response format. The model was entered as shown in equation 17:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Intervention Setting: Pullout } j} + \gamma_2 W_{\text{Intervention Setting: Classroom } j} + \gamma_3 W_{\text{Graphic Organizer } j} + \gamma_4 W_{\text{Retell/summary } j} + u_j + e_j \quad (17).$$

With those predictors entered in the model, a large amount of heterogeneity remained, $Q(96) = 357.92$, $p < .001$, $I^2 = 82.05\%$. With the exception of graphic organizers ($p = .073$), all other variables in the model were significant at $p < .05$ level. Both intervention settings had large, positive with growth from pretest to posttest. Retell/summary had small, positive relationships with growth from pretest to posttest. Specific estimates may be found in Table 10. Together, these findings suggest that retell/summary response format and variables related to intervention were positively and significantly associated with growth.

Research Question 3: Posttests

The goal of RQ 3 was to investigate if intervention outcomes varied based on the particular measure (i.e., response format) used. The outcome of RQ 1 was that response format and intervention type were significant predictors of differences at posttest; however, within the assessment block, retell/summary response formats were the only significant formats. Thus, RQ 3 investigated whether intervention outcomes differed when using a reading comprehension measure with a retell/summary response format.

Retell/summary posttest meta-analysis. A meta-analysis using only effect sizes from measures with retell/summary response formats was conducted. Isolating the sample of effect sizes to this particular format enabled the use of meta-analysis to investigate the association between types of interventions on the specific response format.

The first step was to run an intercept-only meta-analysis (Equation 2) to determine if there was significant heterogeneity in the reading comprehension measures with retell/summary formats posttest effect sizes. Significant heterogeneity would merit further analysis – specifically examining if outcomes vary based on the intervention. An intercept-only meta-analysis was conducted, and findings suggested there was a significant effect of reading interventions on student outcomes on retell/summary response formats at posttest ($p < .001$). The effect was considered large ($\gamma = 0.54$, $SE = 0.17$). In addition findings were associated with significant heterogeneity in effect sizes, $Q\ (21) = 57.30$, $p < .001$, $I^2 = 61.29\%$. The amount of heterogeneity was considered medium. For additional information, these data as well as the intercept-only meta-analytic results for the other response formats, are available in Appendix D.

The significant levels of heterogeneity merited follow-up meta-regression analyses. These analyses may be considered descriptive given the small sample size for these analyses ($k = 13$; $m = 22$). Univariate regressions using intervention types were run. An example is shown Equation 18:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Text Structure Instruction } j} + u_j + e_j \qquad (18).$$

As such, each of the following intervention types from the intervention characteristics block was entered into a univariate regression: graphic organizer, improving background

knowledge, inference instruction, multicomponent, other, self-monitoring, signal words, strategy instruction, technology, text structure instruction, vocabulary component, and writing component.

Of those variables entered into univariate regressions, improving background knowledge and multicomponent interventions were associated with statistically significant and large, positive effects on retell/summary response formats when comparing treatment and control conditions at posttest. No other intervention types were associated with effects.

Both significant variables were entered into a single model. The results for the model are shown in Table 11 in the columns under Post. In addition, the model is represented in Equation 19:

$$g_j = \gamma_0 + \gamma_1 W_{\text{Improving Background Knowledge } j} + \gamma_2 W_{\text{Multicomponent } j} + u_j + e_j \quad (19).$$

Both improving background knowledge ($p < .001$) and multicomponent ($p < .001$) interventions were statistically significant. In addition, both improving background knowledge ($\gamma = 1.35$) multicomponent ($\gamma = 0.86$) interventions were associated with large effects. Furthermore, the model had a nonsignificant amount of heterogeneity, Q (19) = 19.94, $p = .398$. These data are also presented in Table 11 in the columns under Post.

Table 11

Retell/Summary-only Meta-regression with Significant Intervention Variables

	Post [a] (m = 22)			Growth [b] (m = 18)		
Variable	γ	*SE*	p	γ	*SE*	p
Intercept	0.25	0.13	.050	0.97	0.20	<.001

Improving Background Knowledge	1.35	0.32	<.001	-0.59	0.24	.015
Multicomponent	0.86	0.22	<.001	0.49	0.22	.027

Note. m = number of effect sizes; γ = the regression coefficient associated with variable; τ^2 = the estimated between study variance.

[a] $\tau^2 = 0.02$ (*SE* = 0.07), *Q* (19) = 19.94, *p* = .398, $I^2 = 5.98\%$. [b] $\tau^2 = 0$ (*SE* = 0.06), *Q* (15) = 10.73, *p* = .771, $I^2 = 0$

Research Question 3: Growth

Similar to the posttest effect sizes, the only response format that was significant for the growth effect sizes in RQ 2 was retell/summary. As a result, in relation to the growth effect sizes, RQ 3 investigated if student growth within an intervention differs when using a reading measure with a retell/summary response format.

Retell/summary growth meta-analysis. As done with the posttest effects in RQ 3, a meta-analysis was run using only effect sizes from reading comprehension measures with retell/summary formats. An intercept-only meta-analysis (Equation 2) was conducted. In the sample including only growth effect sizes from retell/summary formats, there was a large overall, change in scores from pretest to posttest (γ = 1.08, *SE* = 0.16, *p* < .001). The intercept-only model also revealed significant levels of heterogeneity in the growth effect sizes from retell/summary formats, *Q* (17) = 30.61, *p* = .022, $I^2 = 42.05\%$. The amount of heterogeneity was considered small ($I^2 = 42.05\%$). Given the significant levels of heterogeneity, univariate meta-regressions (e.g., Equation 18) were conducted to explore significant intervention variables associated with differences in outcomes. Of

note, these data are also presented in Appendix D along with the intercept-only meta-analytic results for the other response formats for additional information.

In the univariate regressions, interventions that improved background knowledge and that were multicomponent (i.e., addressed more than one area of reading) were significant. Both variables were included in a final model, and both were statistically significant in the growth analysis. Findings are reported in Table 11 in the columns under, Growth. In addition, findings are described below.

Improving background knowledge had a statistically significant ($p < .001$) and moderate, negative relationship with growth from pretest to posttest in the treatment group ($\gamma = -0.59$, $SE = 0.24$). In contrast, multicomponent interventions had a statistically significant ($p = .04$) and a small, positive relationship with growth from pretest to posttest in the treatment group ($\gamma = 0.49$, $SE = 0.22$). There was not a statistically significant level of between-study heterogeneity present in the model, Q (15) = 10.73, $p = .771$. Similar to RQ 1, there were a limited number of studies included in the analysis (i.e., $k = 10$; $m = 18$). As such, findings should primarily be interpreted as descriptive.

Chapter 5: Discussion

In reading comprehension research and practice, reading comprehension measures are used interchangeably to measure intervention effects (Keenan & Meenan, 2014; Kendeou et al., 2012). Previous researchers (Keenan & Meenan, 2014) and meta-analysts (Collins, 2018; Garcia & Cain, 2014) suggest that different reading comprehension measures account for differing aspects of the construct. Differences in measurement may affect, for example, current knowledge regarding achievement gaps (Collins et al., 2018). The purpose of the current study was to understand if differences in reading comprehension measure's response format was associated with differential intervention outcomes for students.

Findings

The present study was a meta-analysis of 66 published and unpublished research reports and studies conducted within the last twenty years, representing 116 posttest and 104 growth effect sizes. Overall, reading comprehension interventions were associated with small-to-moderate effect size gains in student performance on measures of reading comprehension. The response format of the reading comprehension measure was associated with significant differences in effect size estimates. However, the retell/summary response format was the only format associated with significant differences. Reading comprehension measures with retell/summary formats appeared to be more sensitive to changes in student performance. In contrast, statistically different intervention outcomes were not found for cloze, maze, multiple choice, and sentence

verification response formats. Furthermore, initial evidence suggested intervention outcomes varied within retell/summary response formats. Within retell/summary formats, improving background knowledge and multicomponent interventions were significantly associated with performance at posttest as well as growth from pretest for the treatment group. Additional details regarding findings and their associations to the research literature are described below for each research question. Then implications, limitations, and future directions are explored in the sections that follow.

Research Question 1

RQ 1 was, "In standardized mean difference, as measured through Hedge's g, what is the observed aggregate treatment effect for all reading comprehension interventions delivered to students in elementary school at posttest compared to control?" The purpose of RQ 1 was to broadly understand the effect of reading comprehension interventions on differences between groups at posttest and examine if response format was a significant predictor of variation in effect sizes.

The full sample of available effect sizes in RQ 1 returned an overall weighted mean effect size value of 0.20, which is considered a small effect size difference when using Cohen's (1988) standards. In other words, the treatment groups outperformed the control groups by one-fifth of a standard deviation. These findings broadly relate to previous findings.

Three previous meta-analyses broadly addressed areas similar to RQ 1 by examining the overall effect of reading comprehension interventions (Scamacca, 2015; Shendorevich et al., 2016; Suggate, 2016). Across each meta-analysis, findings were

small, using Cohen's (1988) standards. Although they were all within the small range, the values in the current meta-analyses were not exactly the same. Differences may be due to the population of studies sampled. Suggate (2016) did not restrict based on grade level or publication year; however, all included studies had to examine maintenance effects following the intervention. Shenderovich et al. (2016) sampled a much narrower range of intervention studies, which only included those using peer tutoring. In contrast, Scamacca (2015) sampled a broader range of reading interventions.

Thus, the previously published meta-analytic evidence regarding reading comprehension interventions (Scamacca, 2015; Shendorevich et al., 2016; Suggate, 2016) broadly supports findings from the current meta-analysis. Ultimately, reading comprehension interventions tended to have a positive effect on student outcomes, compared to controls. The magnitude of the effect size may have varied due to differences in study inclusion criteria. However, the overall weighted mean effect size estimate for posttest did not depart from prior findings.

In the current study, there was a small effect of reading comprehension interventions on student outcomes, compared to students who did not receive that intervention. There were moderate-to-large levels of heterogeneity present in the effect sizes. Meta-regression was used to examine characteristics that could explain the moderate-to-large levels of heterogeneity in effect sizes.

Examining response format as a moderator at posttest. The goal of the moderator analysis using meta-regression was to determine if response format was significantly associated with differences between groups at posttest, even after

controlling for purposively selected variables. The purposively selected variables were used to describe meaningful variation in intervention outcomes and serve as controls for any included response format variables.

The final regression model included several variables related to intervention (i.e., continued consultation, technology, graphic organizers, and treatment fidelity), study quality, and measure (i.e., retell/summary formats). Ultimately, the retell/summary response format was significant, even after controlling for the aforementioned purposively selected variables. This suggests that retell/summary response formats were associated with greater effect size differences at posttest. In other words, even after accounting for purposively selected intervention and study quality variables, the treatment group outperformed the control by nearly two-fifths of a standard deviation when monitored using retell/summary formats. This finding was partially corroborated by previous research findings, which are described in the next paragraphs.

In relation to past research findings, Collins' et al. (2018) found significant associations between response format and performance on retell/summary in their meta-analysis. Their findings were similar to those found in the present study. However, they also found significant associations for multiple choice, cloze, and open-ended questions. Their differences in findings may be attributed to differences in the population of interest. Collins et al. (2018) only sampled from studies that included students with reading disabilities *and* students who were typically developing in their reading skills. In addition, their sample spanned grades K-12, did not require the use of an intervention, and excluded special populations from the sample (e.g., studies that focused on English

language learners and students who underperformed in reading and behavior who were otherwise typically developing were excluded). This is different from the population in this study which included only students in elementary school, required an intervention be implemented, and included studies with diverse student populations.

An additional possibility for why findings differed may be due to the research questions investigated in each study. Collins et al. (2018) investigated the role that response format may have on the gap between initial differences in performance between students. In contrast, the present study examined how response format related to differences following the implementation of an intervention. For example, Collins et al. (2018) found that there was a moderate, negative difference between students with reading disabilities and typically developing students, where students with reading disabilities underperformed compared to typically developing students (g = -0.60). Their intercept-only meta-regressions revealed that students with reading disabilities underperformed on average; however, the smallest weighted mean effect size across response formats was in retell/summary ($Range_g$: -1.80 to -0.60). The current meta-analysis found that the largest effect associated with response format at posttest was with retell/summary formats ($Range_g$: -0.14 to 0.44 [Table 7]). Findings from the present study suggest that the largest differences between the control and treatment groups was observed when using retell formats. In this way, the findings from the two studies reveal information about retell formats from different perspectives. Collins et al. (2018) examined preexisting differences at pretest, and this study examined differences following the implementation of an intervention.

Additional reasons for differences between the findings from Collins et al. (2018) and the present study relate to the meta-regression models. To begin, both studies conducted regression models with each response format. Collins and colleagues' (2018) approach differed from the current study, in part given that they ran intercept-only models where an overall weighted mean effect size estimate was calculated using only the data specific to the particular response format. In contrast, the present study entered each variable as a predictor to explain variation in the overall weighted mean effect size estimate across studies (i.e., univariate meta-regressions using response format as the single-predictor). It is possible that the differences in approaches yielded different levels of significance for each response format. However, retrospective power analyses suggested sufficient power to detect a small effect, given the sample size, level of heterogeneity, alpha level of .05, and the level of effect size sought (Harrer et al., 2019).

Continuing with regression methods, the differences in model estimation methods may have also contributed to differences. For example, to calculate their weighted mean effect sizes, Collins et al. (2018) used robust variance estimation to account for multiple effect sizes within studies. The current study used weighted least squares meta-regressions with maximum likelihood estimation and adjusted the weight that each effect size received in order to account for dependence. Robust variance estimation has a number of limitations including a large number of within-study effect sizes ($m \geq 5$ for at least 40 studies; Scammacca et al., 2014). Given the sample of studies included in the current meta-analysis, only one study of the 66 would have met this criteria. In short, both approaches have unique advantages and limitations (Scammacca et al., 2014), and

the analytic approaches could be another factor that contributed to differences beyond sampling, target population, and research questions.

Summary of RQ 1. Overall, there was a small effect of reading comprehension interventions on reading comprehension outcomes. Meta-regressions were used to evaluate the relationship between retell/summary response formats after controlling for purposively selected variables. Retell/summary response formats were associated with significant effects, even after controlling for continued consultation, technology, graphic organizers, treatment fidelity, and evidence of lack of contamination between groups. The overall effect size estimate and the findings that retell/summary was significantly associated with outcomes were corroborated by past research, which highlight the importance of response format (Collins et al., 2018; Scamacca, 2015; Shendorevich et al., 2016; Suggate, 2016). However, past research (Collins et al., 2018) found that other response formats were also significant. These differences in findings may be the result of the studies sampled or the statistical procedures used. Findings from RQ 1 supported the further investigation of the retell/summary response format (i.e., RQ 3). Next, RQ 2 is discussed.

Research Question 2

RQ 2 was, “In standardized mean difference, as measured through Hedge’s g, what is the observed aggregate treatment effect for all reading comprehension interventions delivered to students in elementary school from pretest to posttest?” The purpose of RQ 2 was to understand the effect of reading comprehension interventions on

student growth from pretest to posttest and examine if response format was a significant predictor of variation in effect sizes.

The full sample of available effect sizes in RQ 2 returned an overall weighted mean effect size value of 0.71, which is considered a moderate effect size difference when using Cohen's (1988) standards. In other words, the treatment group improved by nearly three-quarters of standard deviation from pretest to posttest, on average. These findings differ in magnitude and interpretation compared to the findings for RQ 1. This is explored in more detail below.

Exploring differences between growth and posttest effect sizes. In contrast to RQ 1, the effect size estimate for RQ 2 addressed how much a group's score increased or decreased in standard deviation units from before the intervention was implemented to its conclusion. The effect size does not include an equivalent comparison group in its calculation. Therefore, the effects of maturation, the curriculum that all students received, and other supplemental supports delivered by the school for students in the study cannot be ruled out. The result is an effect size that is likely an overestimate of the effect of an intervention because student growth could be related to other factors. Thus, it can be expected that the average effect size would be greater than zero, and the sample of effect sizes would have a negative skew. This can be observed in Figure 2.

In Figure 2, few studies have an effect size value that falls below zero. A score below zero would suggest that from pretest to posttest compared to the groups' original mean score, participants performed worse. This decline would be unlikely considering the passage of time, receiving intervention, participating in school, and possibly receiving

other supplemental supports. One concept to explain the few studies where the growth effect size was zero or below is, regression to the mean (Keith, 2014), which is a statistical phenomenon that describes that scores tend to fall near the mean, and over multiple measurements, high and low scores may occur but most tend to fall close to the mean. Ultimately despite differences in interpretation, the overall weighted mean effect size for growth was another way to estimate the relationship between intervention and measure.

There was no previous meta-analytic research examining the growth of only the treatment group's reading comprehension from pretest to posttest after receiving interventions. However, findings suggested a similar positive trend, relating to positive gains following the implementation of reading comprehension intervention (e.g., Elleman, 2017; Scamacca, 2015; Suggate, 2016). Similar to RQ 1, there were large levels of heterogeneity present in the effect sizes. As a result, meta-regressions were used to examine variables that could help explain the heterogeneity in effect sizes. The results of the moderator analysis are explored next.

Examining response format as a moderator of growth. The same model building procedures used to construct the final model in RQ 1 were used for RQ 2. As such, the final regression model for pretest to posttest growth included purposively selected variables and the retell/summary response format. The goal was to use meta-regression to understand if response format was significantly associated with differences between growth from pretest and posttest, even when controlling for purposively selected variables. Similar to RQ 1, retell/summary was the only response format included in the

final model. However, in contrast to RQ 1, the final regression model for the growth effect sizes only included variables related to measure and intervention because variables within the other blocks were nonsignificant.

Thus, the final regression model consisted of the following variables: graphic organizer, intervention setting – pullout, intervention setting – classroom, and retell/summary response format. Despite the unique nature of the growth effect sizes, the findings for each variable included in the model had little conflict with past research in relation to directionality (e.g., Berkeley et al., 2010; Elleman, 2017; Lee & Tsai, 2017; Swanson et al., 2017). Similar to the posttest analyses, minor differences in the magnitude of estimates in the meta-analysis of growth effect sizes compared to previous meta-analyses could also be due to inclusion criteria and research questions. Ultimately, the retell/summary response format was significantly associated with growth from pretest for posttest.

The results of the model suggested that even after controlling for purposively selected variables, retell/summary response formats had a positive, significant relationship to growth. This particular finding was similar to findings from RQ 1. Specifically, the model estimated effect for retell/summary was quite similar to the value found in RQ 1 (Table 10), which led to a similar interpretation. In other words, even after accounting for purposively selected intervention characteristics, the treatment group grew from pretest to posttest by nearly two-fifths of a standard deviation when intervention outcomes were measured using retell/summary formats.

The relationship between the model estimated effects for retell/summary and findings from past research (i.e., Collins et al., 2018) was similar to those found in RQ 1. In short, the finding that retell/summary formats were significant predictors corroborated past research. However, the lack of other significant response formats was unexpected. The differences in findings could be due to analytic procedures (Scammacca et al., 2014), the research questions, or the sample of included studies. In addition, findings from RQ 2 supported further investigation of retell/summary, specifically if treatment effects for retell/summary differed based on the intervention used (i.e., RQ 3).

Research Question 3

RQ 3 was, "Do observed treatment effects differ based on the specific interventions and measures?" Findings were used to address whether types of reading comprehension measures differentially relate to how intervention outcomes are valued. In both RQ 1 and RQ2, the retell/summary format was a significant predictor of intervention outcomes, even after controlling for purposively selected variables. As such, subgroup analyses were conducted using only effect sizes calculated with a retell/summary measure. All effect sizes associated with a retell/summary format were run with each intervention type as a predictor variable. The purpose was to understand if intervention outcomes varied on retell/summary response formats. Findings for both posttest and growth effect sizes were fairly consistent and are discussed below.

Interventions significantly associated with retell/summary performance. In the case of the posttest analyses, interventions that included multiple components of reading (i.e., fluency, decoding, and/or phonemic awareness) and improving background

knowledge as components of their implementation were associated with significant outcomes compared to the comparison group. Similarly, in the case of the growth analyses, interventions that targeted improving background knowledge and addressing multiple components of reading were also significant outcomes of growth from pretest to posttest. As such, the same variables (i.e., improving background knowledge and multiple components of reading) were observed to be significant in both the growth and posttest analyses. Differences between the two analyses were observed in the magnitude and directionality of the variables' effects.

Differences between the two effect sizes were apparent in the magnitude of effects for multicomponent interventions, and in the magnitude and directionality of effects for improving background knowledge. Specifically, multicomponent interventions were associated with large, positive outcomes in the posttest meta-analysis of the retell/summary data. In contrast, multicomponent interventions in the growth meta-analysis of the retell/summary data were positive, but associated with small-to-moderate effects (Table 11). As such, larger estimates were observed for the posttest group, suggesting multicomponent interventions had greater effects when treatment-to-control at posttest was the standard for comparison. In contrast, multicomponent interventions had a positive but smaller effect when the standard for comparison was growth from pretest to posttest.

In relation to the improving background knowledge, differences in the magnitude and directionality of the effect was observed. In the posttest analysis of retell/summary effect data, improving background knowledge was associated with large, positive

differences in the treatment group compared to control at posttest. In contrast, in the growth analysis, improving background knowledge was associated with a moderate, negative relationship with group growth. The findings from the growth and posttest retell/summary data were quite different and requires additional explanation, which is provided below.

It may appear that the effect of improving background knowledge was contradictory across the two types of effect sizes. In the posttest subgroup analysis for retell/summary, improving background knowledge was associated with a positive estimate. In contrast, in the growth analysis, it was associated with a negative estimate. In both cases, the intercept (see Table 11) is a meaningful variable to consult. In the case of the posttest analysis, the intercept was small ($\gamma = 0.25$), while in the growth analysis it was quite large ($\gamma = 0.97$). In both situations, if a study targeted background knowledge, the overall, mean weighted effect, controlling for the other intervention variable, would be positive. In sum, reasonable conclusions to draw from these findings are that, the differences in the values of the improving background knowledge variables are not as stark when the entire model is taken into account. Now that the results of RQ 3 have been fully reported, their relation to past research is described.

Relating findings from RQ 3 to previous research. Collectively, the findings from the subgroup analyses were aligned with previous research in the area of reading comprehension. The first significant variable was background knowledge. Underperforming readers may lack skills in establishing a coherent mental representation of the text, and may approach the text with low expectations for coherence (Sabatini et

al., 2012). Thus, unskilled readers, who also lack knowledge in a domain, expect the text to be confusing. They may have additional difficulty understanding text without a strong working knowledge of story structure or the content. Improving the background knowledge of students prior to reading is an effective strategy to helping students expect to derive meaning from the text (Elleman, 2017; Joseph, 2015; Kendeou et al., 2016). For example, a common strategy used to support English Language Learners in reading is to build background knowledge by previewing the text (August et al., 2014). As such, the large effects of this intervention component is consistent with what would be expected based on prior research (Sabatini et al., 2012; Kendeou et al., 2012; Kendeou et al., 2016).

The other significant variable was multicomponent interventions. Reading comprehension interventions that were multicomponent, in that they incorporated multiple areas of reading, also had significant and positive associations with outcomes using the posttest and growth data. This finding was also intuitive, given that students who received support in other areas of reading such as fluency and phonemic awareness had higher outcomes in reading comprehension compared to controls (NRP, 2000). In addition, there is a broader range of meta-analytic research highlighting the small-to-large effects that broader reading interventions can have (e.g., Berkeley et al., 2010; Edmonds et al., 2009; Tran et al., 2011). Although the findings for significant variables related to research, there were intervention variables that were unexpectedly nonsignificant in the models explored in RQ 3.

There were several intervention variables that similar arguments could be made for their statistical significance in RQ 3; however, they were not statistically significant predictors of outcomes on the retell/summary response format. Text structure and strategy instructional interventions were two intervention formats that seemed similarly aligned to the retell/summary format. However, they were associated with nonsignificant finding. Graphic organizers are also a way to scaffold the reading comprehension process so that students are able to identify important information within the text (Joseph, 2015). There arguably is a connection between a response format that requires students to organize their ideas and communicate the events of a passage and interventions that aim to support students' representation of their preexisting knowledge as well as their understanding of the text's structure and contents. The lack of statistical significance of these variables, when contextualized with the variables that were significant, call for additional research and investigation. However there are a few possibilities that offer initial insight regarding the outcomes. For each intervention component that was found nonsignificant, it is possible that the issue could be due to a lack of statistical power. It's also possible that the feasibility, or the reliable implementation, of high-quality interventions could be an additional factor for consideration (Fixsen et al., 2005).

Implications

The results of the current study contribute to research on reading comprehension and its measurement. Implications, limitations, and future research are discussed. Findings from the current study have several implications. To begin, researchers may find it useful to include multiple measures of reading comprehension when conducting

studies. When researchers only use one measure of reading comprehension, the measure should be one that can be used consistently (as opposed to interchangeably) and has sufficient reliability and validity information to support its use for the particular research context. However, using multiple measures of reading comprehension, especially when the research questions directly concern reading comprehension, is advisable. A growing line of research suggests that student performance varies, on average, on different measures of reading comprehension (Collins et al., 2018; Kendeou et al., 2012). Furthermore, meta-analytic evidence also suggests that measures correlate differently to the various domains of reading (e.g., decoding, comprehension [Garcia & Cain, 2014; Keenan & Meenan, 2014]). The present study also suggests that in some cases, performance on retell/summary response formats significantly differs, on average, compared to other formats. In sum, to better understand the domain of reading comprehension, multiple measures of varying formats are likely needed. Continuing to conduct research studies using multiple measures of reading comprehension will also allow for future updates to the current meta-analytic sample.

Limitations

As stated in most of the syntheses reviewed in Chapter 2, quantitative studies often fail to report sufficient information on methodology, which results in an inability to code for particular variables. This lack of reported methodological information is a common limitation in educational meta-analyses (Harwell, 2008). In the case of the current meta-analysis, the reporting practices in the primary studies broadly affected the available codes to include in the moderator analyses. This was particularly apparent in the

area of assessment and reader characteristics. Assessment characteristics were rarely provided, leaving few moderators to examine. For example, studies identified and reviewed for this study rarely reported the possible score range or whether the passage remained in view for students to reference. No studies identified and reviewed for this study reported information relevant to fidelity in administering the measures. Likewise, information on race, English proficiency of the sample, and disability status were characteristics of the reader that would have been of interest to include in the moderator analyses, but they were reported at low rates.

In addition, there were studies that met the core criteria for inclusion but did not report necessary information for the calculation of effect sizes. Gate 7 of the inclusion criteria designated that studies needed to provide relevant information to calculate effect sizes. A total of 32 studies were excluded at this gate for not providing or collecting information relevant to calculating effect sizes. Of those studies, nine provided contact information for one of the study authors. Of the seven authors contacted across the nine studies, one author returned the necessary data. As such, there were eight additional studies that met criteria for inclusion, except relevant data were missing and were unable to be retrieved. In sum, this meta-analysis implemented a wide-range and thorough search strategy, but it does not include every published or unpublished study on reading comprehension interventions. Previous meta-analyses in reading comprehension that were reviewed did not report contacting authors for missing data nor the rates of reply, which is commonly unreported factor in meta-analyses in the field of education (Harwell,

2008). This study provides additional detail regarding the collection and inclusion of studies to provide a better understanding of the process and sample of articles included.

There are a variety of methods used to calculate effect sizes (Card, 2012). The current study used Hedge's *g* for treatment and control comparisons at posttest as well as *g* for student growth from pretest to posttest. The use of two methods for effect size calculation is rare, given that it did not occur in any of the 47 meta-analyses reviewed for Chapter 2. The selection of any effect size has to be done thoughtfully, given that different calculations and choices yield different results (Card, 2012; Cooper, Hedges, & Valentine, 2009). There are additional methods for calculating effect sizes (Morris, 2008; Scammacca et al., 2014) that could be beneficial for expanding to other, future analyses to answer different research questions or to expand on the questions that were investigated in the current study.

Another potential limitation is related to the interpretation of effect sizes. Numerous meta-analyses reviewed in Chapter 2 used Cohen's (1988) descriptive categories for effect sizes. Cohen (1988) provided these classifications solely for use as examples, highlighting that the researchers within a discipline should define what constitutes the magnitude of a particular effect (Howell, 2013). As such, the use of this classification system, though prevalent in recent meta-analyses in reading comprehension, may be thought to trivialize clinically significant effects. The labels of small, moderate, and large used to describe effect sizes are only a heuristic, and values may be interpreted differently based on one's familiarity with reading comprehension

research and the contexts in which the data were collected (e.g., elementary school settings).

Future Directions

There are a variety of future directions in research exploring the connections between measures and interventions within reading comprehension. The findings from the current study would help to support research for the questions and projects discussed below.

Investigating connections between assessment and intervention. Future research could investigate how using measures to determine the intervention may relate to improved outcomes. This line of research could expand knowledge regarding the utility of reading comprehension measures for selecting effective interventions for students. A systematic review of using particular reading (comprehension) measures could be beneficial to understanding the relationships between measure and interventions, but also direct applications. Based on the current review, there would be a limited number of studies that used measures to identify which students needed intervention, and even fewer that used measures to identify the particular intervention that students received. A review of that nature would lay the groundwork for experimental studies examining the use of performance on measures to support effective intervention decisions in reading comprehension.

There are hypotheses that may be raised regarding the theoretical linkages between the measures and the interventions. Earlier sections of the discussion explore connections between retell/summary measures and the interventions that build

background knowledge as well incorporate graphic organizers. A future direction could be the implementation of a literature review that delves into the theoretical linkages between specific measures of reading comprehension and available interventions. The literature review could propose a theoretical framework for classifying measures and interventions as distal and proximal. Following the creation of such a framework, it could be applied as a coding scheme using the current meta-analytic sample or another sample. A meta-analysis could then be conducted that investigates the role of the measure's relationship to the intervention. Essentially investigating whether a measure's classification as distal or proximal to the intervention results in heightened treatment effects. Research of this nature would help to distill specific skill mastery measures from general outcome measures within reading comprehension (Fuchs & Deno, 1991).

Investigating differences in retell/summary formats. In the present study, similar evidence was found in both the posttest and growth meta-analyses: reading comprehension measures using retell/summary formats were associated with statistically significant and greater weighted mean effect sizes on average, compared to those observed in measures with other response formats. These reasons could be due to score reliability, true differences in the task and its alignment to intervention, or a combination of the two (Bernfeld et al., 2013; Collins et al., 2018; Fuchs & Fuchs, 1992; Reed et al., 2013). It is possible that the use of retell/summary response formats artificially inflates intervention outcomes due to issues of measurement, or retell/summary response formats accounts for unique variance in the measurement of reading comprehension unaccounted for by measures with other response formats. Ultimately, more research is needed to

further understand the relationship between retell/summary response formats and reading comprehension, and thus how to interpret performance on these measures and the effectiveness of reading comprehension outcomes.

Developers of educational measures may seek to establish a set of standardized materials and procedures for a measure using a retell/summary response format. The measure could be used in a variety of studies to understand its reliability and validity evidence. Building validity evidence for the interpretation and use of the measure for measuring intervention effects would help to understand if the results for retell/summary response formats were associated with unreliability of the measure or true differences in measuring reading comprehension. If the latter is true, a high quality reading comprehension measure with a retell/summary response format may help to support a more comprehensive assessment of student reading comprehension skills.

Future researchers may also use neuroimaging studies as a unique avenue to examine differences in performance on the response formats of reading comprehension measures. Broadly, studies could examine if different regions of the brain are activated when completing measures using varying response formats. As described, reading comprehension is a dynamic and diverse field, investigating the question through a variety of methods would improve the validity evidence supporting general conclusions regarding the importance response formats.

Conclusion

In conclusion, this meta-analysis contributed to a growing body of research syntheses investigating the complexities of reading comprehension measures. It extended

the work to intervention research by exploring reading comprehension response format as a predictor of variability in intervention outcomes. Results in the form of small-to-moderate effect sizes indicated that reading comprehension interventions improve performance for groups of students when compared to controls and to their own scores at pretest. Further analyses led to conclusions that retell/summary response formats were significantly associated with intervention outcomes, even after controlling for intervention and study quality variables. In addition, improving background knowledge and multicomponent interventions were associated with differential outcomes on retell/summary measures. Research in this area will have implications for researchers and practitioners regarding the appropriate uses of reading comprehension measures in monitoring intervention outcomes.

A key conclusion from this study and previous research (Collins et al., 2018) is that students perform differently on the retell/summary response formats. The current study also suggests that interventions work differently on the retell/summary response formats. Use of multiple reading comprehension measures is likely needed to have a holistic understanding of student performance in the domain.

References

*Adams, A. M., Glenberg, A. M., & Restrepo, M. A. (2018). Moved by reading in a spanish-speaking, dual language learner population. *Language, Speech, and Hearing Services in Schools, 49,* 582-594.

Alexander, K. L., Entwisle, D. R., & Kabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record, 103*(5), 760-822.

*Allor, J., & McCathren, R. (2004). The efficacy of an early literacy tutoring program implemented by college students. *Learning Disabilities Research & Practice, 19,* 116-129.

*Amendum, S. J. (2014). Embedded professional development and classroom-based early reading intervention: Early diagnostic reading intervention through coaching. *Reading & Writing Quarterly, 30,* 348-377.

Amendum, S. J., Conradi, K., & Pendleton, M. J. (2016). Interpreting Reading Assessment Data: Moving From Parts to Whole in a Testing Era. *Intervention in School and Clinic*, *51*(5), 284-292.

*Amendum, S. J., Vernon-Feagans, L., & Ginsberg, M. C. (2011). The effectiveness of a technologically facilitated classroom-based early reading intervention: The targeted reading intervention. *The Elementary School Journal*, *112*, 107-131.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, &

Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Assn.

*Anderson, C. M. (2016). *An experimental study of literacy intervention: Teaching foundational reading skills and guided reading* (Doctoral dissertation). Available from ProQuestion Dissertations and Theses database. (UMI No. 10107633)

Araujo, S., Reis, A., Petersson, K. M., & Faisca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *Journal of Educational Psychology, 107,* 868-883.

Ashby-Davis, C. (1985). Cloze and comprehension: A qualitative analysis and critique. Journal of Reading, 28, 585-589.

August, D., McCardle, P., & Shanahan, T. (2014). Developing literacy in English language learners: Findings from a review of the experimental research. *School Psychology Review, 43,* 490-498.

*Baker, L., Dreher, M. J., Shiplet, A. K., Beall, L. C., Voelker, A. N., Garrett, A. J., ... & Finger-Elam, M. (2017). Children's comprehension of informational text: Reading, engaging, and learning. *International Electronic Journal of Elementary Education, 4*, 197-227.

Ball, C. R., & Christ, T. J. (2012). Supporting valid decision making: Uses and misuses of assessment data within the context of RTI. *Psychology in the Schools, 49*(3), 231-244.

Berk, R. A. (1979). The relative merits of item transformations and the cloze procedure for the measurement of reading comprehension. Journal of Reading Behavior, 11, 129-138.

Berkeley, S., Kurz, L., Boykin, A., & Evmenova, A. S. (2015). Improving reading comprehension using digital text: A meta-analysis of interventions. *International Journal for Research in Learning Disabilities, 2*, 18-43.

Berkeley, S., Scruggs, T. E., Mastropieri, M. A. (2010). Reading comprehension instruction for students with learning disabilities, 1995—2006: A meta-analysis. *Remedial & Special Education, 31*, 423-436.

Bernfeld, L. E., Morrison, T. G., Sudweeks, R. R., & Wilcox, B. (2013). Examining reliability of reading comprehension ratings of fifth grade students' oral retellings. Literacy Research and Instruction, 52, 65-86.

*Bohaty, J. J. (2015). *The effects of expository text structure instruction on the reading outcomes of 4th and 5th graders experiencing reading difficulties* (Doctoral Dissertation). Retrieved from https://digitalcommons.unl.edu/cehsdiss/228/

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

*Boulware-Gooden, R., Carreker, S., Thornhill, A., & Joshi, R. M. (2007). Instruction of metacognitive strategies enhances reading comprehension and vocabulary achievement of third-grade students. *The reading teacher, 61*, 70-77.

*Braxton, D. M. (2009a). *The effects of two summarization strategies using expository text on the reading comprehension and summary writing of fourth-and fifth-grade students in an urban, title 1 school* (Doctoral dissertation). Retrieved from https://drum.lib.umd.edu/handle/1903/9918

*Braxton, D. M. (2009b). *Appendix m: Pilot study* (Doctoral dissertation). Retrieved from https://drum.lib.umd.edu/handle/1903/9918

Brown, H. M., Oram-Cardy, J., & Johnson, A. (2013). A meta-analysis of the reading comprehension of individuals on the autism spectrum. *Journal of Autism and Developmental Disorders, 43,* 932-955.

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. Psychology in the Schools, 40, 363-377.

Brown-Chidsey, R., Johnson Jr., P., & Fernstrom, R. (2005). Comparison of grade-level controlled and literature- based maze cbm reading passages. School Psychology Review, 34, 387-394.

Burns, M. K., Petersen-Brown, S., Haegele, K., Rodriguez, M., Schmitt, B., Cooper, M., … Vanderheyden, A.M. (2016). Meta-analysis of academic interventions derived from neuropsychological data. School Psychology Quarterly, *31*, 28-42.

Burns, M. K., Riley-Tillman, T., C., & Rathvon, N. (Eds.). (2017). *Effective school interventions.* New York: The Guilford Press.

Campbell, D.T. & Stanley, J.C. (1963). Experimental and Quasi-Experimental Designs for

Card, N. A. (2012). *Applied meta -analysis for social science research*. New York: The Guilford Press.

Carlisle, J. F. (1989a). The use of the sentence verification technique in diagnostic assessment of listening and reading comprehension. *Learning Disabilities Research, 39,* 159-175.

Carlisle, J. F. (1989b). Diagnosing comprehension deficits through listening and reading. *Annals of Dyslexia*, 39, 159-176.

Carlisle, J. F. (1991). Planning an assessment of listening and reading comprehension. *Topics in Language Disorders*, *12*, 17-31.

Carlisle, J. F. (1999). Free recall as a test of reading comprehension for students with learning disabilities. Learning Disability Quarterly, 22, 11-22.

Carson, K. L. (2017). Reliability and Predictive Validity of Preschool Web-Based Phonological Awareness Assessment for Identifying School-Aged Reading Difficulty. *Communication Disorders Quarterly*.

Chall, J. S. (1983). *Stages of reading development.* New York: McGraw-Hill.

Christ, T. J., & Nelson, P. (2014). Developing and evaluating screening systems: Practical and psychometric considerations. In Kettler, Glover, Albers, & Feeney-Kettler (Eds.), *Universal Screening in Educational Settings: Evidence-Based Decision Making for Schools* (pp. 79-110). Washington, DC: American Psychological Association.

Clark, M. K., Kamhi, A. G., Nippold, M., & Boudreau, D. (2014). Influence of prior knowledge and interest on fourth- and fifth-grade passage comprehension on the qualitative reading inventory—4. Language, Speech & Hearing Services in Schools, 45, 291-301.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd.

Collins, A. A., Lindstrom, E. R., & Compton, D. L. (2018). Comparing students with and without reading difficulties on reading comprehension assessments: A meta-analysis. *Journal of Learning Disabilities, 51,* 108-123.

*Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., ... & Schatschneider, C. (2011). Testing the impact of child characteristics× instruction

interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, *46*, 189-221.

*Connor, C. M., Phillips, B. M., Kim, Y. S. G., Lonigan, C. J., Kaschak, M. P., Crowe, E., ... & Al Otaiba, S. (2018). Examining the efficacy of targeted component interventions on language and literacy for third and fourth graders who are at risk of comprehension difficulties. *Scientific Studies of Reading*, *22*, 462-484.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49,* 560-75.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Sage.

Council for Exceptional Children (2014). CEC: Standards for evidence-based practices in special education. T*eaching Exceptional Children, 46,* 206-212.

Dailor, A. N., & Jacob, S. (2011). Ethically challenging situations reported by school psychologists: Implications for training. *Psychology in Schools, 48,* 619-631.

*Dalton, B., Proctor, C. P., Uccelli, P., Mo, E., & Snow, C. (2011). Designing for diversity: The role of reading strategies and interactive vocabulary in a digital reading environment for fifth-grade monolingual English and bilingual students. *Journal of Literacy Research, 43,* 68-100. doi:10.1177/1086296X10397872

De Santi, R. J., & Sullivan, V. G. (1984). Inter-rater reliability of the cloze reading inventory as a qualitative measure of reading comprehension. Reading Psychology, 5, 37-41.

De Santi, R. J., & Sullivan, V. G. (1985). Reliability of single-rater judgments of semantic and syntactic classifications of cloze test responses. *Journal of Research and Development in Education, 18*, 49-53.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52(3), 219–232.

*Denton, C. A., Anthony, J. L., Parker, R., & Hasbrouck, J. E. (2004). Effects of two tutoring programs on the English reading development of Spanish-English bilingual students. *The Elementary School Journal*, *104*, 289-305.

*Denton, C. A., Tolar, T. D., Fletcher, J. M., Barth, A. E., Vaughn, S., & Francis, D. J. (2013). Effects of tier 3 intervention for students with persistent reading difficulties and characteristics of inadequate responders. *Journal of educational psychology*, *105*, 633.

Dewitz, P., & Dewitz, P. K. (2003). They can read the words, but they can't understand: Refining comprehension assessment. Reading Teacher, 56, 422-435.

*Diebold, T. W. (2011). *Relationship between metacognitive strategy instruction and reading comprehension in at-risk fourth grade students*. Walden University.

*Dixon, O. J. (2007). *Content area readers' theater: The effect on fluency and comprehension* (Doctoral dissertation). University of Houston, TX.

Duffelmeyer, F. A., & Duffelmeyer, B. B. (1987). Main idea questions on informal reading inventories. Reading Teacher, 41, 162-166.

Duke, N. K., & Martin, N. M. (2015). Best practices for comprehension instruction in the elementary classroom. In Parris, S. R., & Headley, K. (Eds.), *Comprehension instruction: Research-based best practices* (pp. 211-223). New York, NY: The Guilford Press.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Scnakenberg, J. W. (2009). *Review of Educational Research, 79,* 262-300.

*Ehri, L. C., Dreyer, L. G., Flugman, B., & Gross, A. (2007). Reading rescue: An effective tutoring intervention model for language-minority students who are struggling readers in first grade. *American Educational Research Journal, 44*, 414-448.

Ehri, L. C., Nunes, S. R., Stah, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the national reading panel's meta-analysis. *Review of Educational Research, 71,* 393-447.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: evidence form the national reading panel's meta-analysis. *Reading Research Quarterly, 36,* 250-287.

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology, 109,* 761-781.

*Engel, K. S. (2018). *Reading comprehension instruction for young students with high functioning autism: forming contextual connections* (Doctoral dissertation). Retrieved from https://academicworks.cuny.edu/gc_etds/2774/

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62,* 497-514.

Field, A. P., & Gillett (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63,* 665-694.

Fixsen, D. L., Naoom, S. F., Blase´, K. A., Friedman, R. M.,& Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication No. 231) Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies. *Educational Psychology Review, 23*, 553-576.

Follmer, J. D. (2018). Executive function and reading comprehension: A meta-analytic review. *Educational Psychologist*, *53*, 42-60.

Fore III, C., Boon, R. T., Burke, M. D., & Martin, C. (2009). Validating curriculum-based measurement for students with emotional and behavioral disorders in middle school. *Assessment for Effective Intervention, 34,* 67-73.

*Frantz, S. O. S. (2001). E*ffectiveness of the infusion of reading component model based remedial reading instruction on the reading achievement of students in learning disabilities and title i remedial reading programs* (Doctoral dissertation). Retrieved from https://elibrary.ru/item.asp?id=5377364

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, 41(1), 93–99.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57,* 488–500.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21,* 1-16.

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities, 45,* 195-203.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. Remedial and Special Education, 9(2), 20–28.

Garcia, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in english. *Review of Educatoinal Research, 84,* 74-111.

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.

*Griffin, L. A. (2010). *Improving reading comprehension instruction: a case study of a professional development initiative* (Doctoral dissertation). Retrieved from https://rucore.libraries.rutgers.edu/rutgers-lib/30069/

*Grogan, S. M. (2014). Reading argumentation and writing: Collaboration and development of a reading comprehension intervention for struggling adolescents (Doctoral dissertation). Available from ProQuestion Dissertations and Theses database. (UMI No. 3645515)

Guthrie, J. T., McRae, A., & Klauda, S. L. (2007). Contributions of concept-oriented reading instruction to knowledge about interventions for motivations in reading. *Educational Psychologist, 42,* 237-250.

*Guthrie, J. T., McRae, A., Coddington, C. S., Lutz Klauda, S., Wigfield, A., & Barbosa, P. (2009). Impacts of comprehensive reading instruction on diverse outcomes of low-and high-achieving readers. *Journal of Learning disabilities, 42,* 195-214.

*Guyne, R. J. H. (2010). The implementation of interventions and strategies for children who struggle with reading utilizing the Read 180 program (Doctoral dissertation). Available from ProQuestion Dissertations and Theses database. (UMI No. 3389735)

Hale, A. D., Hawkins, R. O., Sheeley, W., Reynolds, J. R., Jenkins, S., Schmitt, A. J., & Martin, D. A. (2011). An investigation of silent versus aloud reading comprehension of elementary students using maze assessment procedures. *Psychology in the Schools, 48*, 4-13.

Hale, A. D., Henning, J. B., Hawkins, R. O., Sheeley, W., Shoemaker, L., Reynolds, J. R., & Moch, C. (2011). Reading assessment methods for middle-school students: An investigation of reading comprehension rate and maze accurate response rate. Psychology in the Schools, 48, 28-36.

Hale, A. D., Skinner, C. H., Wilhoit, B., Ciancio, D., & Morrow, J. A. (2012). Variance in broad reading accounted for by measures of reading speed embedded within maze and comprehension rate measures. *Journal of Psychoeducational Assessment, 30*, 539-554.

Haller, E. P., Child, D. A., & Walber, H. J. (1988). Can comprehension be taught: A quantitative synthesis of metacognitive studies. *Educational Researcher, 17,* 5-8.

Hammill, D. D. (2004). What we know about correlates of reading. *Exceptional Children, 70*(4), 453-469.

Hansen, C. L. (1978). Story retelling used with average and learning disabled readers as a measure of reading comprehension. Learning Disability Quarterly, 1, 62-69.

*Haring, C. D. (2013). The effects of coaching on teacher knowledge, teacher practice and reading achievement of at-risk first grade students (Doctoral dissertation). Retrieved from https://repositories.lib.utexas.edu/handle/2152/23148

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019). *Doing Meta-Analysis in R: A Hands-on Guide*. DOI: 10.5281/zenodo.2551803.

Harwell, M., Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *The Journal of Experimental Education*, *76*, 403–430. doi:10.3200/JEXE.76.4.403-430

Hebert, M., Bohaty, J. J., Nelson, J. R., & Brown, J. (2016). The effects of text structure instruction on expository reading comprehension: A meta-analysis. *Journal of Educational Psychology*, *108*, 609-629.

*Hebert, M., Bohaty, J. J., Nelson, J. R., & Lambert, M. C. (2018). Identifying and discriminating expository text structures: An experiment with 4th and 5th grade struggling readers. *Reading and Writing*, *31*, 2115-2145.

Hebert, M., Gillespie, A., & Graham, S. (2013). Comparing effects of different writing activities on reading comprehension: A meta-analysis. *Reading and Writing, 26*, 111-138.

*Hinde, E. R., Popp, S. E. O., Dorn, R. I., Ekiss, G. O., Mater, M., Smith, C. B., & Libbee, M. (2007). The integration of literacy and geography: The Arizona GeoLiteracy program's effect on reading comprehension. *Theory & Research in Social Education*, *35*, 343-365.

Howell, D. C. (2013). *Statistical methods for psychology*. Cengage Learning.

Hudson, R. F., Torgesen, J. K., Lane, H. B., & Turner, S. J. (2012). Relations among reading skills and sub-skills and text-level reading proficiency in developing readers. *Reading and Writing*, *25*, 483–507.

*Jackson, V. (2016). Applying the think-aloud strategy to improve reading comprehension of science content. *Current Issues in Education*, *19*, 1-35.

*Jefferson, R. E., Grant, C. E., & Sander, J. B. (2017). Effects of tier I differentiation and reading intervention on reading fluency, comprehension, and high stakes measures. *Reading Psychology*, *38*, 97-124.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*(4), 582.

*Jennings, C. (2004). The reading together cross-age tutoring program and its effects on the English language proficiency and reading achievement of English language learners (Doctoral dissertation). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.985.4843&rep=rep1&type=pdf

Johnson, E. S., Semmelroth, C., Allison, J., & Fritsch, T. (2013). The technical properties of science content maze passages for middle school students. Assessment for Effective Intervention, 38, 214-223.

*Johnson-Glenberg, M. C. (2000). Training reading comprehension in adequate decoders/poor comprehenders: Verbal versus visual strategies. *Journal of Educational Psychology*, *92*, 772.

Jones, M. B., & Pikulski, J. J. (1979). Cloze for the content area teacher. Reading Research and Instruction, 18, 253-258.

Joseph, L. M. (2015). *Understanding, assessing, and intervening on reading problems* (2nd ed.). Bethesda, MD: National Association of School Psychologists.

Kaldenberg, E. R., Watt, S. J, & Therrien, W. J. (2015). Reading instruction in science for students with learning disabilities: A meta-analysis. *Learning Disability Quarterly, 38,* 160-173.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448.

Kane, M. T . (2006). Validation. In R. L . Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education & Praeger.

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. Journal of Learning Disabilities, 47, 125-135.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: differential dependence on decoding and oral comprehension. Scientific Studies of Reading, 12, 281-300.

Keith, T. Z. (2014). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. Routledge.

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 62-69.

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22,* 354–367.

Kim, J. S., & Quinn, D. M. (2013). The effects of summer reading on low income children's literacy achievement from kindergarten to grade 8: A metaanalysis of classroom and home interventions. *Review of Educational Research, 83,* 386-431.

Kovachy, V. N., Adams, J. N., Tamaresis, J. S., & Feldman, H. M. (2014). Reading abilities in school-aged preterm children: A review and meta-analysis. *Developmental Medicine & Child Neurology, 57,* 410-419.

Kratochwill, T. R., Clements, M. A., Kalymon, K. M. (2007). Response to intervention: Conceptual and methodological issues in implementation. In Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 25-52). New York: Springer Science + Business Media, LLC.

Lai, C. F., Irvin, P. S., Alonzo, J., Park, B. J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM reading comprehension measures: Grade 2* (Technical Report No. 1201). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lee, H. S., & Shu-Fei, T. (2017). Experimental intervention resarch on students with specific poor comprehension: A systematic review of treatment outcomes. *Reading and Writing, 30,* 917-943.

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

Li, H. (2014). The effects of read-aloud accommodations for students with and without disabilities: A meta-analysis. *Educational Measurement: Issues and Practice, 33,* 3-16.

Lietz, P. (2006). Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: A meta-analytic view. *International Education Journal, 7,* 127-149.

Little, S. G., & Akin-Little, A. (Eds.). (2014). *Academic assessment and intervention.* New York: Taylor & Francis.

Magliano, J., Millis, K., Ozuru, Y., & McNamara, D. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), Reading comprehension strategies: Theories, interventions, and technologies, pp.107-136). Mahwah, NJ: Erlbaum.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology*, *47*, 315-355.

*Marshall, H. B. (2017). *The effectiveness of readers' theatre on fluency comprehension and motivation on primary students* (Doctoral dissertation). Retrieved from https://jewlscholar.mtsu.edu/handle/mtsu/5295

*Mason, L. H. (2004). Explicit Self-Regulated Strategy Development Versus Reciprocal Questioning: Effects on Expository Reading Comprehension Among Struggling Readers. *Journal of Educational Psychology*, *96*, 283.

*Mason, L. H., Davison, M. D., Hammer, C. S., Miller, C. A., & Glutting, J. J. (2013). Knowledge, writing, and language outcomes for a reading comprehension and writing intervention. *Reading and Writing*, *26*, 1133-1158.

Mccane-Bowling, S. J., Strait, A. D., Guess, P. E., Wiedo, J. R., & Muncie, E. (2014). The utility of maze accurate response rate in assessing reading comprehension in upper elementary and middle school students. Psychology in the Schools, 51, 789-800.

*McMaster, K. L., van den Broek, P., Espin, C. A., Pinto, V., Janda, B., Lam, E., ... & van Boekel, M. (2015). Developing a reading comprehension intervention: Translating cognitive theory to educational practice. *Contemporary Educational Psychology, 40,* 28-40.

Mcmaster, K. L., Wayman, M. M., & Cao, M. (2006). Monitoring the reading progress of secondary-level english learners: Technical features of oral reading and maze tasks. Assessment for Effective Intervention, 31, 18-31.

*Mead, L. J. (2010). *The effects of using Four Powerful Comprehension Strategies in a gradual release lesson design and learning-style preferences on reading comprehension and self-perception of struggling readers* (Doctoral dissertation). Retrieved from https://repository.wcsu.edu/educationdis/14/

Melby-Lervag, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science, 11,* 512-534.

*Meyer, B. J. F., Wijekumar, K. K., & Lin, Y.-C. (2011). Individualizing a web-based structure strategy intervention for fifth graders' comprehension of nonfiction. *Journal of Educational Psychology*, *103*(1), 140-168. doi:10.1037/a0021606

Meyers, J. (1988). Diagnosis diagnosed: Twenty years after. *Professional School Psychology, 3,* 122-134.

Meyers, J., & Lytle, S. (1986). Assessment of the learning process. *Exceptional Children, 53,* 138-144.

*Miciak, J., Roberts, G., Taylor, W. P., Solis, M., Ahmed, Y., Vaughn, S., & Fletcher, J. M. (2018). The effects of one versus two years of intensive reading intervention implemented with late elementary struggling readers. *Learning Disabilities Research & Practice, 33,* 24-36.

*Miller, C. (2009). *Main idea identification with students with mild intellectual disabilities/specific learning disabilities: A comparison between an explicit and a basal instructional approach* (Doctoral dissertation). Retrieved from http://etd.auburn.edu/handle/10415/1817

Moher, D, Liberati, A. Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Moran, J., Ferdig, R. E., Pearson, P. D., Wardrop, J., & Bomeyer, R. L. (2008). Technology and reading performance in the middle-school grades: A meta-analysis with recommendations for policy and practice. *Journal of Literacy Research, 40,* 6-58.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational research methods, 11,* 364-386.

Muijselaar, M. M. L., Kendeou, P., De Jong, P. F., & Van den Broek, P. W. (2017). What does the cbm-maze test measure. Scientific Studies of Reading, 21, 120-132.

Murphy, K. P., Wilkinson, I. A., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101,* 740-764.

National Center on Intensive Intervention. Academic progress monitoring tools chart. https://charts.intensiveintervention.org/chart/progress-monitoring

National Center on Response to Intervention. Screening tools chart rating system. http://www.rti4success.org/resources/tools-charts/screening-tools-chart/screening-tools-chart-rating-system

National Reading Panel. (2000). *Report of the National Reading Panel Subgroups: Teaching children to read*. National Institute of Child Health and Human Development.

Neddenriep, C. E. (2014). Interventions for developing reading comprehension. In S. G. Akin-Little & A. Little (Eds.), *Academic assessment and intervention* (pp. 219-240). New York, NY: Taylor & Francis.

Neddenriep, C. E., Hale, D. D., Skinner, C. H., Hawkins, R. O., & Winn, B. D. (2007). A preliminary investigation of the concurrent validity of reading comprehension rate: A direct, dynamic measure of reading comprehension. Psychology in the Schools, 44, 373-388.

Neville, D. D., & Searls, E. F. (1991). A meta-analytic review of the effect of sentence-combining on reading comprehension, *Literacy Research and Instruction, 31*, 63-76.

*Newman, L. M. (2007). *The effects of explicit instruction of expository text structure incorporating graphic organizers on the comprehension of third-grade students* (Doctoral dissertation). Retrieved from https://drum.lib.umd.edu/handle/1903/7579

Nilsson, N. L. (2008). A critical analysis of eight informal reading inventories. Reading Teacher, 61, 526-536.

*O'Hara, J. D. (2007). *The influence of supplemental instructional approaches upon the comprehension, metacognitive awareness, and motivation of struggling third-and fourth-grade readers* (Doctoral dissertation). Retrieved from https://drum.lib.umd.edu/handle/1903/6688

O'Reilly T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review, 26,* 403-424.

*Patton, B., Crosby, S., Houchins, D., & Jolivette, K. (2010). The Comparative Effect of Fluency Instruction with and without a Comprehension Strategy for Elementary School Students. *International Journal of Special Education, 25,* 100-112.

Pawlik, K., & Rosenzweig, M. R. (2000). *International handbook of psychology*. Thousand Oaks, CA: Sage Publications.

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., … Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin, 144*, 48-76.

*Phillips, K. M. (2009). *Using graphic organizers to improve at-risk students' reading comprehension of expository text* (Doctoral dissertation). Available from ProQuestion Dissertations and Theses database. (UMI No. 3374412)

Price, K. W., Meisinger, E. B., Louwerse, M. M., & D'Mello, S. K. (2012). Silent reading fluency using underlining: Evidence for an alternative method of assessment. Psychology in the Schools, 49, 606-618.

Rasool, J. M., & Royer, J. M. (1986). Assessment of reading comprehension using the sentence verification technique: Evidence from narrative and descriptive texts. *Journal of Educational Research,* 79, 180-184.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage.

Readance, J. E., & Moore, D. W. (1981). A met-analytic review of the effect of adjunct pictures on reading comprehension. *Psychology in the Schools, 18,* 218-224.

Reed, D. K., Vaughn, S., & Petscher, Y. (2012). The validity of a holistically scored retell protocol for determining the reading comprehension of middle school students. *Learning Disability Quarterly, 35,* 76-89.

Research. Chicago: Rand-McNally.)

*Rhett, T. Y. (2011). *The effectiveness of a reading intervention pull-out program* (Doctoral dissertation, Walden University).

*Ritchey, K. D., Palombo, K., Silverman, R. D., & Speece, D. L. (2017). Effects of an informational text reading comprehension intervention for fifth-grade students. *Learning Disability Quarterly, 40,* 68-80.

*Ritchey, K. D., Silverman, R. D., Montanaro, E. A., Speece, D. L., & Schatschneider, C. (2012). Effects of a tier 2 supplemental reading intervention for at-risk fourth-grade students. *Exceptional Children, 78,* 318-334.

Royer, J. M., & Carlo, M. S. (1991). Assessing the language acquisition progress of limited english proficient students: Problems and a new alternative. *Applied Measurement in Education, 4,* 85-113.

Royer, J. M., Hastings, N. C., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. Journal of Reading Behavior, *11*, 355-363.

Sabatini, J. P., Albro, E. R., & O'Reilly, T. (2012). *Measuring up: Advances in how to assess reading ability.* Plymouth, UK: Rowman & Littlefield Education.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment: In special and inclusive education* (11th ed.). Belmont, CA: Cengage Learning.

*Sanders, S. (2018). *An investigation of the effectiveness of TWA on reading comprehension of students with and at-risk for emotional and behavioral disorders* (Doctoral dissertation). Retrieved from http://krex.k-state.edu/dspace/handle/2097/38663

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

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.

Schulte, A. C. (2016). Prevention and response to intervention: Past, present, and future. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (pp. 59-71). New York, NY, US: Springer Science + Business Media.

Sencibaugh, J. M. (2007). Meta-analysis of reading comprehension interventions for students with learning disabilities: strategies and implications. *Reading Improvement, 44,* 6-22.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). Improving Reading Comprehension in Kindergarten through 3rd Grade: IES Practice Guide. NCEE 2010-4038. *What Works Clearinghouse*.

Shapiro, E. S., Fritschmann, N. S., Thomas, L. B., Hughes, C. L., & Mcdougal, J. (2014). Concurrent and predictive validity of reading retell as a brief measure of reading comprehension for narrative text. Reading Psychology, 35, 644-665.

Shenderovich, Y., Thurston, A., & Miller, S. (2016). Cross-age tutoring in kindergarten and elementary school settings: A systematic review and meta-analysis. *International Journal of Educational Research, 76,* 190–210. https://doi.org/10.1016/j.ijer.2015.03.007

*Simmons, D., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Willson, V., ... & Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness*, *3*, 121-156.

Skinner, C. H., Williams, J. L., Morrow, J. A., Hale, A. D., Neddenriep, C. E., & Hawkins, R. O. (2009). The validity of reading comprehension rate: Reading speed comprehension and comprehension rates. Psychology in the Schools, 46, 1036-1047.

*Smith, A., & Feng, J. (2018, October). *Literature circle and gifted students: Boosting reading motivation and performance*. Paper presented at the Annual Meeting of the Georgia Educational Research Association, Macon, GA.

Smith, N., & Zinc, A. (1977). A cloze-based investigation of reading comprehension as a composite of subskills. Journal of Reading Behavior, 9, 395-398.

Snow, A. B., Morris, D., & Perney, J. (2018). Evaluating the effectiveness of a state-mandated benchmark reading assessment: mClass Reading 3D (Text Reading and Comprehension). *Reading Psychology*, *39*(4), 303-334.

*Solari, E. J., Denton, C. A., Petscher, Y., & Haring, C. (2018). Examining the effects and feasibility of a teacher-implemented Tier 1 and Tier 2 intervention in word reading, fluency, and comprehension. *Journal of Research on Educational Effectiveness*, *11*, 163-191.

Spargo, E. (1989). *Timed readings* (3rd ed.). Providence, RI: Jamestown.

Spector, J. E. (2005). How reliable are informal reading inventories. *Psychology in the Schools*, *42*, 593–603.

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. School Psychology Review, 39, 258-276.

Spencer, M., & Wagner, R. K. (2017). The comprehension problems for second-language learners with poor reading comprehension despite adequate decoding: A meta-analysis. *Journal of Research in Reading, 40*, 199-217.

Spencer, M., & Wagner, R. K. (2018). The comprehension problems of children with poor reading comprehension despite adequate decoding: A meta-analysis. *Review of Educational Research*, 40, 366-400.

Stevenson, N. A., Reed, D. K., & Tighe, E. L. (2016). Examining potential bias in screening measures for middle school students by special education and low socioeconomic status subgroups. Psychology in the Schools, 53, 533-547.

Suggate, S. P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of Learning Disabilities, 49*, 77-96.

Swanson, E., Stevens, E. A., Scammacca, N. K., Capin, P., Stewart, A. A., & Austin, C. R. (2017). The impact of tier 1 reading instruction on reading outcomes for students in grades 4-12: A meta-analysis. *Reading and Writing: An Interdisciplinary Journal, 30*, 1639-1665.

Swanson, H. L. (1999). Reading research for students with LD: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities, 32*, 504-532.

Taylor, J. B. (1983). Influence of speech variety on teachers' evaluation of reading comprehension. Journal of Educational Psychology, 75, 662-667.

Thorndike, R. M., & Thorndike-Christ, T. (2011). *Measurement and evaluation in psychology and education* (8th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Tindal, G., & Parker, R. (1989). Development of written retell as a curriculum-based measure in secondary programs. School Psychology Review, 18, 328-343.

Tolar, T. D., Barth, A. E., Francis, D. J., Fletcher, J. M., Stuebing, K. K., & Vaughn, S. (2012). Psychometric properties of maze tasks in middle school students. Assessment for Effective Intervention, 37, 131-146.

*Tong, F., Irby, B. J., Lara-Alecio, R., & Koch, J. (2014). Integrating literacy and science for English language learners: From learning-to-read to reading-to-learn. *The Journal of Educational Research*, *107*, 410-426.

*Tong, F., Irby, B. J., Lara-Alecio, R., & Mathes, P. G. (2008). English and Spanish acquisition by Hispanic second graders in developmental bilingual programs: A 3-year longitudinal randomized study. *Hispanic Journal of Behavioral Sciences, 30,* 500-529.

*Tong, F., Lara-Alecio, R., Irby, B. J., & Mathes, P. G. (2011). The effects of an instructional intervention on dual language development among first-grade Hispanic English-learning boys and girls: A two-year longitudinal study. *The Journal of Educational Research, 104,* 87-99.

*Trainin, G., Hayden, H. E., Wilson, K., & Erickson, J. (2016). Examining the impact of QuickReads' technology and print formats on fluency, comprehension, and vocabulary development for elementary students. *Journal of Research on Educational Effectiveness, 9,* 93-116.

Tran, L., Sanchez, T., Arellano, B., & Swanson, H. L. (2011). A meta-analysis of the rti literature for children at risk for reading disabilities. *Journal of Learning Disabilities, 44,*283-295.

Trezek B. J., & Mayer, C. (2015). Using an informal reading inventory to differentiate instruction: Case studies of three deaf learners. *American Annals of the Deaf, 160,* 289-302.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). 2019. *Reading Assessments*. https://www.nationsreportcard.gov/reading_2017/#/nation/achievement?grade=4

Van den Broek, P. (1988). The effects of causal relations and hierarchical position on the importance of story statements. *Journal of Memory and Language, 27,* 1–22.

van den Broek, P., Beker, K., & Oudega, M. (2015). Inference generation in text comprehension: automatic and strategic processes in the construction of a mental representation. In O'Brien, E. J., Cook, A. E., & Lorch, R. F., Jr. (Eds.), *Inferences during reading* (pp. 94 -121). Cambridge, UK: Cambridge University Press.

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J. S., Butler, J., White, M. J., & Lorch, E. P. (2005). Assessment of comprehension abilities in young children. In S. Stahl & S. Paris (eds.), Children's Reading Comprehension and Assessment, (pp.107-130). Mahwah, NJ: Erlbaum.

*Vaughn, S., Chard, D. J., Bryant, D. P., Coleman, M., Tyler, B. J., Linan-Thompson, S., & Kouzekanani, K. (2000). Fluency and comprehension interventions for third-grade students. *Remedial and Special Education, 21,* 325-335.

*Vaughn, S., Roberts, G. J., Miciak, J., Taylor, P., & Fletcher, J. M. (2019). Efficacy of a word- and text-based intervention for students with significant reading difficulties. *Journal of learning disabilities, 52,* 31-44.

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

*Vaughn, S., Wanzek, J., Murray, C. S., Scammacca, N., Linan-Thompson, S., & Woodruff, A. L. (2009). Response to early reading intervention examining higher and lower responders. *Exceptional Children, 75*, 165-183.

*Vernon-Feagans, L., Kainz, K., Hedrick, A., Ginsberg, M., & Amendum, S. (2010). The Targeted Reading Intervention: A Classroom Teacher Professional Development Program to Promote Effective Teaching for Struggling Readers in Kindergarten and First Grade. *Society for Research on Educational Effectiveness*.

Walcyzk, J. J. (1990). Relation among error detection, sentence verification, and low-level reading skills of fourth graders. *Journal of Educational Psychology, 82*, 491-497.

Walcyzk, J. J., & Royer, J. M. (1989). A program for constructing SVT tests: An alternative way of assessing text comprehension. *Behavior Research Methods, Instruments & Computers, 21*, 369-370.

*Wanzek, J., & Roberts, G. (2012). Reading interventions with varying instructional emphases for fourth graders with reading difficulties. *Learning Disability Quarterly, 35*, 90-101.

*Wanzek, J., Petscher, Y., Al Otaiba, S., Kent, S. C., Schatschneider, C., Haynes, M., ... & Jones, F. G. (2016). Examining the average and local effects of a standardized treatment for fourth graders with reading difficulties. *Journal of Research on Educational Effectiveness, 9*, 45-66.

*Wanzek, J., Petscher, Y., Otaiba, S. A., Rivas, B. K., Jones, F. G., Kent, S. C., ... & Mehta, P. (2017). Effects of a year long supplemental reading intervention for students with reading difficulties in fourth grade. *Journal of Educational Psychology, 109*, 1-17.

What Works Clearinghouse. (2019). Retrieved from https://ies.ed.gov/ncee/wwc/WhatWeDo

*Wijekumar, K. K., Meyer, B. J., & Lei, P. (2013). High-fidelity implementation of web-based intelligent tutoring system improves fourth and fifth graders content area reading comprehension. *Computers & Education, 68,* 366-379.

*Wijekumar, K., Meyer, B. J., Lei, P. W., Lin, Y. C., Johnson, L. A., Spielvogel, J. A., ... & Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade readers. *Journal of Research on Educational Effectiveness, 7,* 331-357.

*Williams, J. P., Kao, J. C., Pao, L. S., Ordynans, J. G., Atkins, J. G., Cheng, R., & DeBonis, D. (2016). Close analysis of texts with structure (CATS): An intervention to teach reading comprehension to at-risk second graders. *Journal of Educational Psychology*, *108,* 1061.

Willingham, D. T. (2007). Ask the cognitive scientist: The usefulness of brief instruction in reading comprehension strategies. *American Educator, 30,* 39–45.

Wood, S. G., Moxley, J. H., Tighe, E. L., & Wagner, R. K. (2018). Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities: A meta-analysis. *Journal of Learning Disabilities, 51,* 73-84.

*Zipke, M., Ehri, L. C., & Cairns, H. S. (2009). Using semantic ambiguity instruction to improve third graders metalinguistic awareness and reading comprehension: An experimental study. *Reading Research Quarterly*, *44,* 300–321. doi:10.1598/RRQ.44.3.4

Appendix A

Initial Contact Email: Amendum, Braxton, Connor, Denton, Mason, Ritchey, Vaughn, Wanzek, Wijekumar, Williams

Hello Dr. ____,

My name is Calvary Diggs, and I am a graduate student at the University of Minnesota. I am contacting you because of your work in reading comprehension research.

I am currently completing my dissertation, which is a meta-analysis investigating if reading comprehension intervention outcomes vary based on the selected assessment. I am seeking unpublished data in order to reduce publication bias; I wanted to know if you have any relevant studies that have not yet been published. If so, would you be able to complete the excel included below for any relevant studies.

Of note, I am specifically interested in studies that were conducted in the United States in K-12 school settings, with assessment and intervention methods conducted in English. The intervention(s) has to be related to reading comprehension (i.e., …) and the measure(s) must fit at least one of the following formats (i.e.,). In addition, the study must have included a pretest and a posttest as well as a control group.

Very best,

Calvary Diggs

University of Minnesota

Treatment					Control					Additional Treatments				
Intervention Type (dropdown)					Intervention Type (dropdown)					Intervention Type (dropdown)				
Measure Type (dropdown)														
N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}
Measure Type (dropdown)														
N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}

Follow-up Email(s):

Hello Dr. ____,

I wanted to send a reminder email regarding your ability to provide any unpublished data for inclusion in my meta-analysis on how assessment choice may moderate intervention outcomes in reading comprehension assessment and intervention work. Additional details are available below. Please let me know if you have any questions or concerns.

Appendix B

Initial Contact Email: Abe, Crowe, Register, Torgensen, Trainin, Williams, Wijekumar

Hello Dr. ____,

My name is Calvary Diggs, and I am a graduate student at the University of Minnesota. I am contacting you because of your article titled, _____.

I am currently completing my dissertation, which is a meta-analysis investigating if reading comprehension intervention outcomes vary based on the selected assessment. I am seeking additional information regarding effect sizes so that your study may be included in my analyses. I wanted to know if you could provide the ___ (mean, standard deviation)___ for the ___(treatment, control)___ group. If so, would you be able to complete the excel included below?

Treatment					Control					Additional Treatments				
Intervention Type (dropdown)					Intervention Type (dropdown)					Intervention Type (dropdown)				
Measure Type (dropdown)														
N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}
Measure Type (dropdown)														
N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}	N	M_{pre}	SD_{pre}	M_{post}	SD_{post}

Very best,

Calvary Diggs

University of Minnesota

Follow-up Email(s):

Hello Dr. ____,

I wanted to send a reminder email regarding your ability to provide missing information for calculating effect sizes for your study titled, _____. I would like to include your study in my meta-analysis on how assessment choice may moderate intervention outcomes in reading comprehension assessment and intervention work. Additional details are available below. Please let me know if you have any questions or concerns.

Very best,

Calvary Diggs

University of Minnesota

Appendix C

Displayed equations from Chapter 4 are provided below. Lines were used to support readability

$$w_{adjusted} = \frac{1}{\Delta_j * m_j} \quad (1)$$

$$g_j = \gamma_0 + u_j + e_j, \quad (2)$$

$$g_j = \gamma_0 + \gamma_1 W_{1j} + \gamma_2 W_{2j} + \ldots + \gamma_s W_{sj} + u_j + e_j \, . \quad (3)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Control Condition: Treated Control } j} + \gamma_2 W_{\text{Control Condition: Other } j} + \gamma_3 W_{\text{Level of randomization: Student } j} + \gamma_4 W_{\text{Level of randomization: Classroom } j} + \gamma_5 W_{\text{Level of randomization: School } j} + \gamma_6 W_{\text{Matching } j} + \gamma_7 W_{\text{Number of Measures } j} + \gamma_8 W_{\text{Publication Type: Dissertation } j} + \gamma_9 W_{\text{Publication Type: Other } j} + \gamma_{10} W_{\text{Publication Year } j} + \gamma_{11} W_{\text{Quasi-experiment } j} + u_j + e_j \quad (4)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Control Condition: Treated Control } j} + \gamma_2 W_{\text{Number of Measures } j} + u_j + e_j \quad (5)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Continued Consultation } j} + \gamma_2 W_{\text{Implementer: Researcher } j} + \gamma_3 W_{\text{Implementer: Classroom Teacher } j} + \gamma_4 W_{\text{Implementer: Other } j} + \gamma_4 W_{\text{Treatment Fidelity } j} + u_j + e_j \quad (6)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{ Graphic Organizer } j} + \gamma_2 W_{\text{ Improving Background Knowledge } j} + \gamma_3 W_{\text{ Inference Instruction } j} + \gamma_4 W_{\text{ Multicomponent } j} + \gamma_5 W_{\text{ Other } j} + \gamma_6 W_{\text{ Self-monitoring } j} + \gamma_7 W_{\text{ Signal Words } j} + \gamma_8 W_{\text{ Strategy Instruction } j} + \gamma_9 W_{\text{ Technology } j} + \gamma_{10} W_{\text{ Text Structure Instruction } j} + \gamma_{11} W_{\text{ Vocabulary Component } j} + \gamma_{11} W_{\text{ Writing Component } j} + u_j + e_j \quad (7)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{ Continued Consultation } j} + \gamma_2 W_{\text{ Treatment } \textit{Fidelity } j} + \gamma_3 W_{\text{ Graphic Organizer } j} + \gamma_4 W_{\text{ Improving Background Knowledge } j} + \gamma_5 W_{\text{ Technology }} + u_j + e_j \quad (8)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{ Administration Size } j} + \gamma_2 W_{\text{ Administrator } j} + \gamma_3 W_{\text{ Standardized Measure } j} + u_j + e_j \quad (9)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{ Cloze } j} + u_j + e_j \quad (10)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Summary or Retell } j} + u_j + e_j \quad (11)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Grade (Minimum) } j} + \gamma_2 W_{\text{Grade (Maximum) } j} + u_j + e_j \quad (12)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Quality 6.3 } j} + \gamma_2 W_{\text{Quality 6.8 } j} + \gamma_3 W_{\text{Quality 7.3 } j} + u_j + e_j \quad (13)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Continued Consultation } j} + \gamma_2 W_{\text{Graphic Organizer } j} + \gamma_3 W_{\text{N Measures } j} + \gamma_4 W_{\text{Summary/Retell } j} + \gamma_5 W_{\text{Technology } j} + \gamma_6 W_{\text{Quality 6.3 } j} + u_j + e_j \quad (14)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Continued Consultation } j} + \gamma_2 W_{\text{Implementer: Researcher } j} + \gamma_3 W_{\text{Implementer: Classroom Teacher } j} + \gamma_4 W_{\text{Implementer: Other } j} + \gamma_4 W_{\text{Treatment Fidelity } j} + \gamma_5 W_{\text{Intervention Setting: Classroom } j} + \gamma_6 W_{\text{Intervention Setting: Pullout } j} + \gamma_7 W_{\text{Intervention Setting: Tutorial } j} + u_j + e_j \quad (15)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Intervention Setting: Classroom } j} + \gamma_2 W_{\text{Intervention Setting: Pullout } j} + \gamma_3 W_{\text{Graphic Organizer } j} + u_j + e_j \quad (16)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Intervention Setting: Pullout } j} + \gamma_2 W_{\text{Intervention Setting: Classroom } j} + \gamma_3 W_{\text{Graphic Organizer } j} + \gamma_4 W_{\text{Summary/Retell } j} + u_j + e_j \quad (17)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Text Structure Instruction } j} + u_j + e_j \quad (18)$$

$$g_j = \gamma_0 + \gamma_1 W_{\text{Text Structure Instruction } j} + \gamma_1 W_{\text{Text Structure Instruction } j} + u_j + e_j \quad (19)$$

APPENDIX D

Data from intercept-only meta-analyses are provided below. Each meta-analysis was conducted using effect sizes from one specific response format. Results for effect sizes from posttest analyses are presented in the rows titled, 'Posttest.' Results for effect sizes from growth analyses are presented in the rows titled, 'Growth.'

Intercept-only regressions for all response formats using posttest effect sizes and growth effect sizes

Response Format	Analysis	*k*	*m*	*g*	*SE*	*p*	*95% CI*	τ^2	*Q*	I^2
Cloze	Posttest	26	34	0.10	0.09	.262	[-0.08, 0.28]	.16	<.001	70.87%
	Growth	24	32	0.70	0.11	<.001	[0.48, 0.92]	.30	<.001	84.01%
Maze	Posttest	4	4	0.16	0.17	.362	[-0.18, 0.49]	0	.823	0%
	Growth	4	4	0.52	0.35	.143	[-0.17, 1.20]	.34	.007	71.35%
Multiple-Choice	Posttest	34	41	0.12	0.06	.026	[0.02, 0.23]	.06	<.001	71.06%
	Growth	28	34	0.60	0.09	<.001	[0.42, 0.78]	.20	<.001	88.19%
Open-ended	Posttest	11	14	0.20	0.21	.332	[-0.21, 0.62]	.43	<.001	78.44%
	Growth	9	11	0.79	0.29	.006	[0.23, 1.35]	.67	<.001	83.88%

Retell/Summary	Posttest	13	22	0.54	0.17	.001	[0.21, 0.87]	.32	<.001	61.29%
	Growth	10	18	1.08	0.16	<.001	[0.77, 1.38]	.37	.022	42.05%
True/False	Posttest	1	2	0.48	0.67	.472	[-0.83, 1.79]	.66	.055	72.76%
	Growth	1	2	1.03	0.42	.015	[0.20, 1.85]	.16	.183	43.71%

Note. k = number of studies; m = number of effect sizes; g = Hedge's g mean effect size; τ^2 = the estimated between study variance; Q = statistical test of heterogeneity between effect sizes; I^2 = estimate of the percentage of heterogeneity between study effect sizes compared to the total variability in effect sizes.