Relations between CBM (Oral Reading and Maze) and Reading Comprehension on State Achievement Tests: A Meta-Analysis

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Jaehyun Shin

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Kristen McMaster, Adviser

September, 2017

Acknowledgement

There were many people who helped me complete this dissertation and my doctoral program, which is a big milestone in my life. First, I would like to express my deepest gratitude to my adviser, Dr. Kristen McMaster, who served as my great mentor throughout my doctoral program. Dr. McMaster has given me support, guidance, and confidence whenever I need to complete each stage of my graduate school experience. I am grateful that I had my role model as a scholar.

I would like to thank my dissertation committee members, Dr. Scott McConnell, Dr. Michael Harwell, and Dr. Lori Helmen, who encouraged and inspired me to pursue my dissertation study. I am deeply grateful for their continuous support, guidance, and valuable feedback throughout this project.

Finally, I would like to say a special "thank you" to Dr. Dong-Il Kim, who guided me to the field of special education and helped me grow into a researcher. Your support and encouragement enabled me to pursue my Ph.D. at the University of Minnesota.

Dedication

I dedicate my dissertation to my mother, for your unconditional love and sacrifice.

I also dedicate this dissertation to God. In God’s divine guidance, I could take a single step at a time.

Abstract

The purpose of this study was to examine the validity of two widely used Curriculum-Based Measurement (CBM) in reading – oral reading and maze task – in relation to reading comprehension on state tests using a meta-analysis. A total of 61 studies (132 correlations) were identified across Grades 1 to 10. A random-effects meta-analysis was conducted to estimate the average correlations between the two CBMs and reading comprehension on state tests, and to analyze the effects of potential moderating variables (characteristics of study, students, CBM, and state tests). Results revealed that the average correlation for oral reading was significantly larger than that for maze when all grade levels were included together in the analysis. When grade levels were separated, the difference between average correlations was only at the higher grades (Grades 4-10), favoring oral reading. In terms of correlations by grade level, oral reading and maze showed a similar pattern; that is, correlations were comparable across elementary grades, but decreased for secondary grades. In addition to the type of CBM and grade level differences, type of publication, development type of state tests (commercial versus state-developed), and time interval between CBM and state tests were significant sources of variance in correlations. Implications for research and educational practice are discussed highlighting the somewhat different conclusions from previous literature, especially regarding the use of CBM for older students.

Table of Contents

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

With a nationwide emphasis on accountability standards for meeting academic success in reading, assessing student reading achievement and growth has become increasingly important (Decker, Hixson, Shaw, & Johnson, 2014; Gibbons & Casey, 2012; Wanzek, Roberts, Linan-Thompson, Vaughn, Woodruff, & Murray, 2010). Because states and districts are accountable for students' year-end achievement status, researchers and policymakers recommend that educators use benchmark screenings to identify students at risk for not meeting state standards in reading (Nese, Park, Tindal, & Alonzo, 2011; Wanzek et al., 2010). Further, screening is an essential component in multi-tiered instructional systems of support (e.g., response to intervention; RTI) for identifying students who may not respond to core instruction and are in need of additional intervention (Nese et al., 2011).

To identify students at risk for failing state reading standards as measured by high-stakes state or district-administered tests, many schools and districts have implemented Curriculum-Based Measurement (CBM, Deno, 1985) as a benchmark system (Nese et al., 2011; Silberglitt & Hintze, 2005; Wood, 2006). CBM is a standardized measurement procedure for assessing students' academic status and growth in the basic skill areas of reading, math, spelling, and written expression (Deno, 1985; 2003). CBM has been widely used for progress monitoring over the past three decades (Stecker, Fuchs, & Fuchs, 2005) and recently has been commonly used as part of universal screening to identify students in need of additional monitoring and more intensified intervention

(Fuchs, 2004). CBM is an evidence-based and useful approach to assess students' reading status and progress, and most-applied systematic assessment of academic performance (Nese et al., 2011).

CBM Reading Tasks as a Valid Indicator of Reading Proficiency

Research has shown that CBM in reading is a valid indicator of general reading proficiency as measured by standardized tests (Ardoin, Witt, Suldo, Connell, Koenig, Resetar et al., 2004; Wayman et al., 2007), based largely on criterion-validity evidence. Two types of CBM reading tasks – oral reading and maze – have received the most attention as potential screening tools in research and school settings (McMaster, Wayman, & Cao, 2006; Wayman et al., 2007). Oral reading requires students to read a passage aloud for 1 min; the score is the number of words read correctly. On maze, students silently read a passage in which every seventh word is deleted, and select the correct word among three response choices. Students usually read for 1 to 3 min; the number of correct selections is scored.

The majority of research on CBM has focused on oral reading rather than maze in relation to general reading proficiency (Fuchs, Fuchs, & Compton, 2004; Stecker et al., 2005). Accumulated studies have indicated that CBM oral reading can be considered a proxy of general reading proficiency as measured by standardized achievement tests, with correlations reported to range from r = .60 to .75 (Nese et al., 2011) or over .80 (Wayman et al., 2007). Because CBM oral reading has been found to correlate with broad measures of reading proficiency, including reading comprehension, it is considered to be more than just a measure of fluent decoding (Fuchs et al., 2001; Wayman et al., 2007).

Nevertheless, correlations between scores on CBM oral reading and standardized reading achievement tests tend to decline as a function of the increase in grade level (e.g., Jenkins & Jewell, 1993; Silberglitt, Burns, Madyun, & Lail, 2006), indicating that oral reading may not be a valid indicator for older students. In addition, some researchers have criticized oral reading for emphasizing decoding speed rather than comprehension in that a student could read the passage quickly without comprehending the meaning of the text (Reidel, 2007). For this reason, oral reading may appear to overestimate the reading competency of children whose primary reading difficulties relate to comprehension problems rather than decoding or fluency (Hamilton & Shinn, 2003).

The maze task, on the other hand, has received less attention to date in screening research than has oral reading (Stecker et al., 2005; Wayman et al., 2007). As researchers extend the scope of participants (e.g., to older students) and continue to question whether oral reading is the best proxy of reading proficiency, especially reading comprehension, maze has become a focus of interest (Wayman et al., 2007). Evidence indicates that maze performance is associated with accurate and efficient word reading (i.e., fluency) and with skills supporting the construction of a mental representation of text including vocabulary knowledge and inference making (Kendeou et al., 2012). In addition, for teachers, maze often is more easily accepted as a measure of reading comprehension than is oral reading (Graney et al., 2010; Wayman et al., 2007).

Compared to the decreasing trend in the relation between oral reading and criterion measures as grade level increases, maze has shown a consistent trend across grade levels with moderate correlations (Graney et al., 2010). validity of oral reading and maze

(Jenkins & Jewell, 1993), correlations between oral reading and criterion measures (i.e., standardized reading measures) decreased from *r* =.80 in Grades 2 through 4 to *r* =.60 and *r* =.70 in Grades 5 and 6. Correlations for maze remained fairly consistent across Grades 2 to 5, ranging from *r* =.65 to *r* =.75. For this reason, maze has been considered to be more acceptable for older students, such as those in intermediate (Grades 4-6) or secondary students (Wayman et al., 2007).

Need for Screening Tools to Identify Students with Reading Comprehension Difficulties

Although an extensive research base supports the relation between CBM reading tasks and performance on standardized tests of reading (Wayman et al., 2007), much of this work has targeted students in the early elementary grades (Grades 1-3) who typically struggle with basic reading skills, such as decoding or reading fluency. By contrast, less work has been conducted with students in the intermediate grades who may have difficulty with reading comprehension. In addition, as the primary emphasis on reading instruction shifts from word decoding and fluency to comprehension and learning from text, the choice of screening tools may also need to shift (Graney et al., 2010). The reason why screening tools for reading comprehension are needed is that students with specific reading comprehension difficulties have unique characteristics compared to their peers whose primary difficulties are at the word decoding or fluency level (Cutting & Scarborough, 2006; Perfetti et al., 2005). Further, a failure to identify reading comprehension problems early is likely to contribute to the pervasive reading difficulties that become more and more difficult to remediate (Elleman et al., 2011).

Reading comprehension is a multidimensional, complex process that requires readers to decode efficiently, draw on vocabulary and background knowledge, remember what they have read, and analyze text (Cain, Oakhill, & Bryant, 2000; Kendeou, Papadopoulos, & Spanoudis, 2012; Lesaux & Kieffer, 2010). Based on the framework known as the "*Simple View of Reading*" (Gough & Tunmer, 1986), reading comprehension is a product of word recognition (i.e., ability to read isolated words quickly and accurately) and linguistic comprehension (i.e., ability to use linguistic knowledge to derive meaning from texts). Thus, the simple view of reading divides reading comprehension into two separate processes.

Basic reading skills (e.g., fluent decoding) are critical to succeed in reading comprehension for early elementary students by freeing cognitive resources required for processing of meaning (Cutting & Scarborough, 2006; LaBerge & Samuels, 1974), and those skills have shown to predict performance on reading comprehension tests (Catts, Herrera, Nielsen, & Bridges, 2015; Kendeou et al., 2012). However, mastering those skills may not guarantee success in reading comprehension (Cutting & Scarborough, 2006). For example, students might correctly decode words but not understand the text (Cutting & Scarborough, 2006; Perfetti et al., 2005).

Other language skills, such as syntax or vocabulary, have shown to account for additional variance in predicting reading comprehension, while the role of basic reading skills may decrease over time (Catts, Adlof, & Weismer, 2006). Those language skills are known to be associated with specific reading comprehension deficits (i.e., poor comprehenders; Cutting & Scarborough, 2006), particularly in the upper elementary school grades. Poor comprehenders' primary difficulty is comprehension in spite of having adequate decoding or fluency skills (Cain & Oakhill, 2007; Nation & Snowling, 1997). Poor comprehenders are also known to lack sufficient working memory and higher-order processes such as inference making (Cain et al., 2000).

With respect to the difference between poor comprehenders and those whose main difficulties are at the word level, research indicates that reading comprehension difficulties could occur in the absence of word recognition problems (Bishop & Snowling, 2004). Based on cognitive models of reading comprehension (Perfetti et al., 2005), their difficulties are text-based (e.g., making connections within the text, deriving word meaning from the text) or at the situation model level (e.g., using background knowledge to make inferences), not at the surface level such as decoding (Fletcher et al., 2007). In this respect, poor comprehenders are predicted to have relatively higher standing on CBM oral reading than on other reading comprehension measures (Hamilton & Shinn, 2003; Wayman et al., 2007). Thus, CBM oral reading might not always be sufficient for detecting reading comprehension difficulties, particularly for older elementary or secondary school students (Elleman et al., 2011).

How to assess reading comprehension and its related indicators (e.g., core cognitive processes) has been controversial (Kendeou et al., 2012; van den Broek, Rapp, & Kendeou, 2005; Wayman et al., 2007). This controversy is partly because current measures of reading comprehension may not tap on the same linguistic or cognitive processes (Cutting & Scarborough, 2006) or because the measures may lack technical adequacy, such as validity (RAND Reading Study Group, 2002). Researchers have recently focused on screening tools for reading comprehension difficulties, asking significant questions about what the tools are actually measuring (Elleman et al., 2011). Given that the insensitivity of reading measures to detect reading comprehension difficulties might hinder early identification of students with reading comprehension difficulties (Elleman et al., 2011), more investigation is warranted to find the most adequate screening measures across different grades.

Further, timely screening of students with reading comprehension difficulties is critical because comprehension difficulties may not obviously emerge and are often overlooked at primary grades (Compton, Fuchs, Fuchs, Elleman, & Gilbert, 2008; Elleman et al., 2010). Compton and colleagues (2008) revealed that students can be identified as having ‘late-emerging’ reading disabilities--especially related to comprehension problems--after grades 4-5. In addition, those students are likely to be missed from identification after they receive word recognition or fluency interventions (Compton et al., 2008). This difficulty is partly because responsiveness to early word-level interventions does not necessarily predict later reading comprehension difficulties (Compton et al., 2008). These findings corroborate the importance of timely and accurate screening of students with potential reading comprehension difficulties.

Problem Statement

Thus far, the research base on the criterion validity of CBM reading tasks in relation to specific reading comprehension skills, which become increasingly important as students get older, provides only preliminary results. In addition, despite that CBM maze appears to be more of a reading comprehension measure than oral reading does (e.g., Fuchs, Fuchs, & Maxwell, 1988; Shin, Deno, & Espin, 2001; Wiley & Deno, 2005), the previous literature does not strongly suggest that maze is superior to oral reading in assessing reading comprehension (Wayman et al., 2007).

In sum, given that it is still not clear how closely CBM reading tasks relate to reading comprehension, more empirical evidence is needed to support whether CBM oral reading and maze are valid screening measures for identifying specific reading comprehension difficulties across grades (Elleman et al., 2011; Graney et al., 2010). Moreover, additional evidence should be provided about which CBM tasks better predict reading comprehension on standardized achievement tests across grade levels. In doing so, practitioners may gain a better understanding of which screening tools to use to identify students who struggle specifically with reading comprehension. In using the most appropriate screening tools, they may be more likely to provide appropriate interventions that target students' specific needs in a timely manner (Graney et al., 2010).

Study Purpose and Research Questions

Examining the criterion validity of CBM reading tasks in relation to state tests of reading comprehension has practical significance given that those state tests are frequently used under school accountability systems (Graney et al., 2010). In addition, relevant factors (i.e., moderators) should be taken into account when considering the relations between CBM reading tasks and standardized achievement tests because those factors may influence variability in these relations (Wayman et al., 2007; Wood, 2006). According to previous literature, those factors include, but are not limited to, grade level differences, participant characteristics (e.g., English language learner [ELL] status), characteristics of criterion measures (e.g., state or nationally normed, response format), and the time interval between the administration of CBM and criterion measures (e.g., Kranzler et al., 1999; Silberglitt et al., 2006; Wood, 2006; Yeo, 2009).

To provide integrative conclusions regarding the relations between CBM (oral reading and maze) and reading comprehension along with the effects of potential moderating factors, a meta-analytic review provides a useful approach to statistically synthesize a large collection of findings from an existing research base (Glass, McGaw, & Smith, 1981). Recently, three meta-analytic reviews on this topic were published (Reschly, Busch, Betts, Deno, & Long, 2009; Yeo, 2009, 2010). However, Reschly et al. (2009) and Yeo (2009) focused on CBM oral reading only. Yeo (2010) focused on both oral reading and maze, but examined the relations between CBM and overall reading achievement on state tests, rather than on reading comprehension specifically.

Therefore, the current study will make a unique contribution to the literature beyond what the extant reviews have made by providing a comprehensive examination of the research base to support the utility of CBM oral reading and maze as indicators of reading comprehension as measured by state achievement tests. Findings from this meta-analysis will provide converging evidence on the overall strength of these relations, along with factors that account for variance in these relations (e.g., different population characteristics, the nature of state achievement tests) across studies. Thus, these findings will shed light on practitioners' understanding about whether CBM oral reading and maze can be valid screening tools of reading comprehension across variant conditions or education settings.

Specific research questions are as follows:

RQ 1. What are the estimated average correlations between CBM tasks (oral reading and maze) and reading comprehension on state achievement tests, and do they vary by (a) type of task (oral reading vs. maze) and (b) grade level (primary, intermediate, secondary)?

- Hypothesis 1. For primary grades (Grades 1-3), the estimated average correlation between oral reading and reading comprehension will be larger than that between maze and reading comprehension.
- Hypothesis 2. For intermediate and secondary grades (Grades 4-12), estimated average correlation between maze and reading comprehension will be larger than that between oral reading and reading comprehension.

RQ 2. To what extent are the relations between CBM tasks (oral reading and maze) and reading comprehension on state achievement tests influenced by potential moderating factors?

CHAPTER II

LITERATURE REVIEW

The purpose of this literature review was to summarize previous qualitative and quantitative syntheses on the relations between CBM reading tasks (oral reading or maze) and reading comprehension on criterion measures for Grades 1-12. Another purpose was to identify potential factors that might influence the relations between CBM and criterion measures of reading comprehension. By doing so, I sought to determine what further examination is needed to address the relations between CBM reading tasks and reading comprehension on state achievement tests.

Method

Literature Search Procedure

Review studies on the relations between CBM oral reading and maze and reading comprehension on state achievement tests were located via the following steps. First, electronic databases (ERIC, PsycINFO, and Google Scholar) were used to search for relevant literature syntheses. Descriptors for the database searches included combinations of the following keywords: “curriculum-based measurement,” “reading aloud,” “oral reading fluency,” and “maze” with “reading comprehension,” “reading achievement,” and “achievement tests.” Second, ancestral searches were conducted by examining reference lists of identified articles. Third, recent volumes of the following peer-reviewed journals that commonly deal with identification of and intervention for students with reading difficulties were hand searched: *Exceptional Children, Journal of Learning*

Disabilities, Learning Disabilities Research and Practice, Journal of Special Education, Learning Disability Quarterly, School Psychology Review, School Psychology Quarterly, and Remedial and Special Education.

Inclusion Criteria

To qualify as relevant to the present review of literature, literature syntheses were examined to determine whether they met the following inclusion criteria. First, review studies had to focus on examining the relations between CBM (oral reading and/or maze) and criterion reading measures. Second, review studies had to be either qualitative or quantitative reviews (i.e., meta-analysis) on the relations. Third, participants in the studies reviewed had to be students from Grade 1 to 12 with and without reading difficulties or disabilities. Fourth, review studies had to be conducted in the U.S. and had to be written in English.

Results

Base on the inclusion criteria, four previous review studies (one qualitative synthesis and three meta-analyses) were identified. Below, I summarize those four reviews with a focus on the validity of CBM oral reading and maze.

Qualitative Synthesis

Wayman et al. (2007) qualitatively synthesized the research base on CBM in reading published since the time of Marston's review in 1989. Marston's review, published in *Curriculum-Based Measurement: Assessing Special Children* (Shinn, 1989), summarized the initial work on CBM in reading; Wayman and colleagues aimed to update the field since this initial work. Marston (1989) reviewed fourteen CBM studies on word identification and oral reading, with a focus on technical adequacy to provide support for the use of CBM reading as indicators of overall reading proficiency. Correlations between oral reading and criterion measures ranged from *r* = .57 to .90. In addition, Marston's (1989) review demonstrated that oral reading was correlated with teacher judgment and with other measures of reading comprehension.

Wayman et al. (2007) focused on the technical adequacy (reliability and validity) of the three most commonly-used CBM measures, namely oral reading (reading aloud), maze, and word identification. A total of 64 studies were included in the review. Participants were students from Grades 1 through 8, including students in general education, receiving ELL services, or receiving special education. The CBM oral reading task was used in over 80% of the studies; maze, on the other hand, was used in only 10 studies, indicating that many fewer studies had been conducted for maze at the time of Wayman et al.'s (2007) review.

Correlations between oral reading and state achievement tests (Iowa, Michigan, Minnesota, and Washington) of reading for Grades 1-5 ranged from *r* = .49 to *r* = .81. A relatively wider range of correlations was reported between oral reading and commercial criterion reading measures (*r* = .21 to .93). Correlations between maze and state achievement tests (Iowa, Minnesota) of reading ranged from *r* = .46. to *r* = .73, and correlations with commercial criterion reading measures ranged from *r* = .50 to *r* = .76.

Wayman and colleagues' (2007) review confirmed earlier research findings that demonstrated a strong association between oral reading and overall reading proficiency.

In addition, oral reading was found to be a better indicator of reading comprehension than other reading comprehension measures, such as cloze, retell, and question-answering (Fuchs, Fuchs, & Maxwell, 1988). The authors interpreted this result to mean that oral reading was more than just a measure of fluent decoding. Rather, findings provide empirical support for the theorized relation between oral reading and reading comprehension.

The authors, however, highlighted that correlations between oral reading and criterion measures of reading tended to decrease as students' grade levels increased, although correlations remained moderate (*r*s = .50 to .70) to strong (*r*s > .70) across elementary grades. On the other hand, correlations between maze and criterion measures of reading tended to remain fairly stable across the grades. Therefore, the authors suggested that oral reading might be the best CBM task for primary grades, both oral reading and maze may be appropriate for intermediate grades, and maze may be the best choice for secondary grade students (Wayman et al., 2007).

The authors also pointed out that more research on the validity of CBM reading tasks is needed for older students (e.g., Grades 6-12) and students with diverse backgrounds, given that the overwhelming majority of CBM validity research had been done with oral reading for students in Grades 2-5, and evidence indicated that oral reading may overestimate performance of ELL or African American students (Wayman et al., 2007). In addition, the authors suggested that other factors (e.g., passage characteristics) needed to be examined because those factors may influence the relations between CBM and criterion measures of reading.

Wayman and colleagues (2007) conducted a very thorough review on CBM reading tasks with a focus on technical adequacy. However, the synthesis was qualitative rather than quantitative, preventing a precise estimation of the overall strength of the relations between two CBM tasks (i.e., oral reading and maze) and criterion measures of reading and the effects of potential moderating variables. Further, given that most studies in this review reported the correlations between CBM reading and general reading proficiency measured by norm-referenced tests, it was difficult to estimate the relations with reading comprehension specifically.

Quantitative Synthesis

Next, three quantitative, meta-analytic reviews were summarized as follows.

Reschly et al. (2009) summarized the relations between CBM oral reading and standardized measures of reading achievement for students in Grades 1-6 using a meta-analysis. A total of 41 studies from 1980 to 2008 were included for the meta-analysis. These studies included a diverse group of participants, such as ELL, students receiving special education, and students eligible for free or reduced lunch. Potential moderating variables (e.g., source of test, administration format, length of time between CBM and criterion tests, and reading subtest type) were examined to determine whether those variables influenced the magnitude of the relations between oral reading and standardized measures of reading achievement.

Results indicated that the correlation across all studies was *r* = .68. Specifically, correlations ranged from *r* = .35 to .91 (with nationally-normed tests) and from *r* = 43 to .81 (with state-developed tests). In addition, significant moderating effects were found for source of criterion test (state versus nationally-normed test), administration format (group or individual), time interval between CBM and criterion test, and reading subtest type. Specifically, the correlation with national tests was significantly higher than the correlation with state tests. Also, the correlation of individually administered tests was significantly higher than the correlation of group administered tests. Further, the correlations significantly decreased when the time interval between CBM and criterion tests increased. Finally, oral reading was more significantly and highly correlated with word identification than with other reading skills, such as vocabulary or comprehension. However, there were no significant differences in correlations across grade levels.

Reschly et al.'s (2009) study demonstrated that oral reading functions as a good indicator of criterion tests of reading, with a higher correlation with nationally-normed tests than with state-specific tests. The authors explained that one reason for this finding was that nationally-normed tests were designed to gauge general reading achievement, whereas state tests were developed to assess grade-level standards set by each state and were likely to have varying difficulty levels and quality (Reschly et al., 2009). Although a lower correlation was found for state tests, the association with state tests was moderately strong. Another notable finding of the study was that the relations between oral reading and criterion tests of reading did not significantly decrease as students progressed from grade 1 through 6. The authors reserved their opinion on this issue given that there was insufficient data to investigate the correlations as a function of grade levels (Reschly et al., 2009).

This study, however, could not examine the relations between oral reading and criterion measures of reading as a function of student demographic characteristics (e.g., ethnicity, students with special education services, ELL) because many studies did not report demographic information specific to the participants of the studies, whereas others reported information for the participants but not specific to grade level. As the authors mentioned, examining the effects of student demographic characteristics (e.g., different ethnic or language backgrounds) is important to establish the validity of inferences drawn from any tests.

Another meta-analysis, conducted by Yeo (2008), also examined the relations between CBM oral reading and various kinds of reading comprehension measures. The comprehension measures included CBM retelling, CBM maze, commercial criterion measures, and state achievement tests (e.g., Florida Comprehensive Assessment Test, California Achievement Test). In addition, this study examined moderators' effects on the variability of CBM oral reading in relation to reading comprehension measures. A total of 55 studies with 250 correlations were identified for the meta-analysis.

The estimated mean correlation between CBM oral reading and various reading comprehension tests was r = .75, with a range of r = .71 to .79. Based on Cohen's (1992) criteria for interpreting effect sizes, the magnitude of the overall relations was large. Given significant variability between studies, conditional meta-analyses were conducted including potential moderators. These moderators included the characteristics of participants (grade level, proportion of female, free or reduced lunch, ELL, special education in the sample), CBM (number of passages, type of passage, administrator of

CBM), reading comprehension tests (response format, type of test [criterion or norm-referenced], time interval between oral reading and reading comprehension tests), and study (e.g., type of publication, year of publication).

Meta-analysis results revealed that the strength of the relations between oral reading and reading comprehension tests significantly differed by the proportions of students with disabilities, response format of reading comprehension tests, and the time interval between CBM and reading comprehension tests. Specifically, studies with a high proportion of students receiving special education showed higher correlations between oral reading and reading comprehension tests. In terms of the effect of response format, retelling was negatively associated with correlations. The time interval between CBM and reading comprehension tests was negatively related to correlations, indicating that correlations were higher as the time intervals between CBM and reading comprehension tests were shorter.

Yeo (2008) conducted a quantitative review of the studies on the relations between CBM oral reading and various kinds of reading comprehension tests using a multi-level meta-analysis. Results of this study provided evidence for the strong association between oral reading and reading comprehension tests. Moreover, this study revealed that the relations could be influenced by potential moderatos, such as participants' characteristics or criterion measures' (reading comprehension tests) characteristics. However, this study examined correlations of only CBM oral reading in relation to reading comprehension. In addition, given that this study treated CBM maze as a criterion measure of reading comprehension, the direct comparison between oral reading and maze was not examined.

The last meta-analysis, conducted by Yeo (2010), examined two CBM reading measures (oral reading and maze) and students' performance on state achievement tests in reading. Twenty-seven studies met the inclusion criteria. Results showed an overall correlation between CBM and state achievement tests in reading of *r* = .689, and correlations significantly differed from study to study. Given this heterogeneity, a series of moderator analysis were conducted. Moderators included characteristics of participants (e.g., grade level, proportion of ELL), CBM (type of CBM [oral reading or maze], type of passage [school curriculum or standardized]), state achievement tests (e.g., response type, time interval between CBM and state tests), and study (e.g., type of publication).

The moderator analyses revealed that study sample size, proportion of ELL students, proportion of students with disabilities, type of CBM, and time interval between CBM and state achievement tests were significant moderators that influenced the relations between CBM and state achievement tests in reading. Specifically, correlations were positively related to the study sample size and negatively related to the proportion of ELL and special education students. Regarding type of CBM, oral reading appeared to have a stronger relation with state tests in reading than did maze. In addition, as expected, longer time intervals between CBM and state tests administrations resulted in lower correlations.

Results of this study, however, should be interpreted with caution in some ways. First, grade level was not a significant moderator, which means correlations between CBM and state tests did not significantly differ by grade levels. However, most grade levels included in this meta-analysis ranged from Grades 1 through 5. Only one study in this meta-analysis included Grades 6 and above. Therefore, more studies that include higher grade levels (e.g., secondary grades) are needed to examine the effect of grade level on the relations between CBM and state tests. In addition, although the type of CBM (oral reading or maze) was a significant moderator in this study, only three studies on maze were included in the meta-analysis. Further research is needed to investigate whether there is a significant effect of type of CBM across grade levels.

Discussion

The purpose of this literature review was to summarize previous syntheses on CBM in reading. By doing so, this literature review sought to answer what further research is needed to determine the relations between CBM (oral reading and maze) and reading comprehension measured by state achievement tests.

Evidence of the Validity of CBM Oral Reading and Maze

The four review studies indicated that CBM oral reading and maze could be valid indicators of reading proficiency on criterion measures. Specifically, oral reading appeared to predict students' overall reading performance better than did maze for early elementary grades, whereas maze appeared to be a better indicator for intermediate and secondary level students (Wayman et al., 2007). In addition, when other conditions were held constant, oral reading had a significantly stronger relation with students' performance on state reading achievement tests across grade levels than did maze (Yeo, 2010).

In terms of moderator effects, several variables were found to significantly influence the relations between CBM and criterion measures of reading. Those moderators included sample size, proportions of students with disabilities, proportion of ELL students, type of CBM, time interval between CBM and criterion tests, source of criterion tests (state versus nationally-normed test), administration format (group or individual) and response format of criterion tests (Reschly et al., 2009; Yeo, 2008; Yeo, 2010).

The three meta-analyses, however, did not fully reveal the relations between two CBMs (oral reading and maze) and reading comprehension on criterion tests. First, Reschly et al. (2009) and Yeo (2008) focused on CBM oral reading only; therefore, it was not possible to compare the validity of oral reading with the validity of maze. Second, Yeo (2010) focused on both oral reading and maze, but examined the relations between CBM and overall reading achievement on state tests, rather than reading comprehension specifically. Third, the number of studies on maze was too small (only three studies) to provide converging evidence for maze as a valid indicator of reading comprehension across grade levels. Therefore, further research is needed to provide more empirical evidence for the relations between CBM oral reading and maze and reading comprehension.

Potential Moderating Factors

Despite the limitations above, the four previous reviews summarized in this chapter provide clear implications for further research, especially for potential moderating factors that might influence the relations between CBM and reading comprehension measured by state achievement tests. In addition to the four reviews, other individual studies that were identified from those reviews were discussed together to provide converging results regarding the potential moderating factors.

First, regarding the effects of participant characteristics on the relation between CBM and other reading measures, the comparison between students with and without disabilities (Yeo, 2010) and between ELL and non-ELL (Wiley & Deno, 2005; Yeo, 2010) appeared to influence the validity of CBM in relation to criterion measures of reading. Specifically, Wiley and Deno (2005) revealed that the validity of CBM might differ for groups with different language backgrounds by showing that maze had a strong relation with the Minnesota state test in reading for non-ELL only. One possible explanation is that ELL students' fluency may not be sufficiently developed to decode automatically. Ethnicity and SES also seemed relevant to the variation in the relations. Pearce and Gayle (2008) showed that the contribution of SES and ethnicity in predicting the Dakota state test was statistically significant in Grades 3 and 5. In addition, Kranzler et al. (1999) showed that performance on oral reading appeared to overestimate reading comprehension of African American students and underestimate that of Caucasians. Regarding the effect of SES, Paleologos and Brabham (2011) found significant correlations between oral reading and SAT-10 for high-income students with proficient fluency levels but not low-income students with proficient fluency skills, suggesting that early vocabulary instruction might be necessary especially for low-SES students.

Second, the nature of state tests of reading (e.g., state-specific or nationally developed, response format) might influence the relations (Reschly et al., 2009; Wayman et al., 2007; Yeo, 2010). As Wanzek and colleagues (2010) pointed out, different predictive validity of oral reading can be produced according to the type (nationally-normed or state-normed) of states' high-stake tests. In addition, Stage and Jacobson (2001) reported relatively low correlations between oral reading and the Washington state achievement test as compared to those in other studies with the same grade (e.g., Kranzler et al., 1999; Marcotte & Hintze, 2009; Wood, 2006). One possible explanation for this result is that the state-normed test may involve differential constructs or measure different aspects in comparison to other state tests of reading comprehension (Baker et al., 2008).

A third factor that possibly influences the relations is the time interval between CBM and criterion measure administration (Reschly et al., 2009; Yeo, 2008; 2010). Roehrig et al. (2008) also showed that correlations between oral reading and two standardized tests scores tended to decrease as the difference in time between administrations increased. Many studies have indicated this tendency, however, some studies have demonstrated inconsistent results. For example, Shapiro et al. (2006) study showed that the time interval between two administrations did not influence the relations for Grade 4 (*r* = .68 in fall, *r* = .69 in winter, and *r* = .66 in spring). As for the time interval effect, it appears that too many other sources of variations exist across studies to simply conclude that longer time intervals yield a lower correlation (Yeo, 2010).

A fourth source of variation in the relations would be type of CBM scores which is used to predict reading comprehension scores on state tests. Although the meta-analyses did not explore this effect, Baker et al. (2008) showed that the use of slope values over time added predictive power to oral reading as a potential indicator of reading comprehension as measured by a large-scale federal reading proficiency test. Stage and Jacobson (2001) examined fourth grade oral reading scores as well as slope in predicting students' performance on Washington Assessment of Student Learning (Taylor, 2000). However, the correlation with slope over time (across the school year) predicted WASL scores less well (r = .35) than level of performances on oral reading (r = .50 and r = .51 for fall and winter/spring, respectively). Given that few studies have examined whether slope values on oral reading add predictive power for estimating performance on state tests of reading comprehension, more investigation appears to be needed (Baker et al., 2008; Wanzek et al., 2010).

Conclusion

This literature review sought to synthesize the existing literature on the validity of CBM oral reading and maze. Qualitative and quantitative syntheses indicated that oral reading and maze task can be used as valid indicators of overall reading proficiency. Still, several questions remain unanswered. First, few studies on maze were included in previous syntheses as compared to oral reading. Second, more evidence is needed about the validity of oral reading and maze as valid indicators of reading comprehension, rather than overall reading proficiency. Thus, further meta-analytic syntheses are needed to provide converging empirical evidence of the relations between CBM (oral reading and maze) and reading comprehension, along with potential effects of various moderating factors on these relations.

CHAPTER III

METHOD

Literature Search Procedure

The studies reviewed in this meta-analysis were located via three steps. First, electronic databases (ERIC, PsycINFO, and Google Scholar) were used to search for the relevant literature in peer-reviewed journals, dissertations, and technical reports on the relation between scores on CBM and state achievement tests in reading comprehension. The search was completed in November of 2016. Descriptors for the database searches included combinations of the following keywords: “curriculum-based measurement,” “reading aloud,” “oral reading fluency,” and “maze” with “reading comprehension,” “reading achievement,” and “achievement tests.” Websites of commercially-available CBM systems (AIMSweb, DIBELS, easyCBM, and FAST) and Research Institute for Progress Monitoring (RIPM) were also searched for technical reports. Second, ancestral searches were conducted by examining reference lists of identified articles and recent literature reviews on the relation between CBM and achievement tests in reading comprehension (Reschly et al., 2009; Wayman et al., 2007; Yeo, 2009; Yeo, 2010). Third, recent volumes of the following peer reviewed journals (published between May and November of 2016) that commonly deal with participants with reading difficulties were hand searched: *Exceptional Children, Journal of Learning Disabilities, Learning Disabilities Research and Practice, Journal of Special Education, Learning Disability Quarterly, Assessment for Effective Intervention* and *Remedial and Special Education*.

Inclusion Criteria

To qualify as relevant for the present meta-analysis, studies were examined to determine whether they met the following inclusion criteria. First, studies had to examine the relation between CBM (oral reading and/or maze) and state achievement tests of reading comprehension and include concurrent or predictive validity information. Regarding reading comprehension, studies were included if either a specific reading comprehension component (subtest) was clearly stated as part of the state achievement tests or the state tests appeared to reflect reading comprehension although reading comprehension was not specifically stated (e.g., students read six passages that covered a broad range of topics and completed 56 multiple choice questions related to those passages; Oregon Statewide Assessment in Reading; Oregon Department of Education, 2000). In addition, studies that used nationally-normed standardized tests of reading comprehension were included because some states use commercially prepared standardized tests, such as the Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2005) and Stanford Achievement Test (SAT; Harcourt Educational Measurement, 2003). Studies that only examined relations between standardized reading achievement test scores (e.g., relation between fluency and comprehension subtest of Gray Oral Reading Test, Wiederholt & Bryant, 2001) or between CBM scores (e.g., relationship between reading aloud and maze) were excluded.

Second, studies in which CBM was administered before or concurrently with standardized reading comprehension tests were included, to determine concurrent or predictive validity. Third, studies had to provide sufficient quantitative information (e.g., correlations, *t* statistics, *F* statistics) so that correlation coefficients could be calculated.

In addition, studies that focused primarily on interventions and that used pre/post or longitudinal designs were included if they provided sufficient data for examining correlations between scores on CBM and achievement tests of reading comprehension. Fourth, participants had to be students from Grade 1 to 12 with or without reading difficulties or disabilities. Fifth, only studies that used correlational or group experimental designs were included; studies that used single-subject designs or that provided only qualitative data were excluded. Sixth, only studies conducted in the U.S. and written in English were included.

Coding System

Once all studies that met inclusion criteria were identified, a coding system was developed. In addition, descriptors (e.g., participants, type of CBM used, state achievement tests used, study level) and detailed information about each descriptor were recorded. Table 1 includes the specific coding guide.

Table 1

Coding of Study Descriptors

Category	Definition of Descriptors
Participants	
Grade level	The grade level of sample students
Grade range	Combined grade levels (primary: Grade 1-3; intermediate: Grades 4-6; secondary: Grades 7-10)
Sample size	The number of participants
Female %	The percentage of female participants
ELL %	The percentage of ELL participants
Special education %	The percentage of participants receiving special education
Free reduced lunch (FRL) %	The percentage of participants receiving FRL
White %	The percentage of White participants

Category	Definition of Descriptors
CBM	
Number of CBM passages	The number of passages used to obtain one score
Type of CBM task	The type of CBM task, recorded as (1) oral reading (2) maze
Development type of CBM passage	The development type of passage, recorded as (1) standardized (2) researcher-developed (3) other (4) NA
Who administered CBM	Administrator of CBM, recorded as (1) researchers (including graduate research assistants (2) school personnel (3) NA
State achievement test	
Name of tests	The name of state tests used
Development type of state tests	The development type of state tests, recorded as (1) commercially developed (2) state developed (3) other (4) NA
Norm- or criterion-referenced state tests	Recorded as (1) norm-referenced (2) criterion-referenced (3) NA
Response formats	The response formats, recorded as (1) multiple choice (2) open-ended questions (3) combination of multiple choice and open-ended questions (4) other (5) NA
Time interval between CBM and state achievement test	The time interval (in months) between CBM and state achievement tests was recorded
Study level	
Year of publication	Publication year for each study was recorded
Type of publication	The type of publication, recorded as (1) journal article (2) dissertation (3) technical report
Correlation coefficient	The correlation coefficient between CBM and state tests was recorded

In the Participants category, the “grade range” code was added to “grade level” for examining which type of CBM (oral reading or maze) better predicts reading comprehension on state tests for primary (Grade 1-3), intermediate (Grade 4-6), and secondary levels. Given that researchers have suggested that oral reading is a better indicator of reading comprehension for primary grades and maze for intermediate and secondary grades (e.g., Wayman et al., 2007), analyses based on the grade range were conducted to provide empirical evidence of these relations. Regarding the time interval, if the authors did not clearly mention the time interval between administrations, I calculated the approximate time (in months) between CBM and state achievement tests administration. For example, if CBM was administered in early September and state tests in early December, the time interval was coded as '3.'

Search and Coding Reliability

Interrater agreement was established for the search of studies based on the inclusion criteria, as well as for coding the variables within the studies. For the search interrater agreement, approximately 20% of the studies identified for coding were randomly selected and screened by one doctoral student in educational psychology. Interrater agreement was 84%; all disagreements were resolved through discussion. For coding agreement, I served as the primary coder, and one doctoral student in curriculum and instruction served as a second independent coder. The second coder was trained on how to use the coding sheet, definitions of study variables, and recording effect sizes, before coding independently. Guided practice and feedback was provided as needed as part of training. To establish coding reliability, approximately 20% of all studies was selected at random and coded by both raters, and the results of coding was compared. Coding reliability result was 94.1%. All disagreements were discussed and resolved by consensus.

Meta-Analytic Approach

A fixed-effects model assumes that the observed effect sizes provide information about the results of the particular studies chosen by a researcher, but does not provide information about the generalizability of effect sizes to other studies (Hedges & Vevea, 1998; Rosenthal et al., 2006). In other words, the purpose of using a fixed-effects model is to draw a conditional inference only about the studies included in the meta-analysis (Hedges & Vevea, 1998). On the other hand, random-effects models assume that effect sizes vary across studies (Borenstein, Hedges, Higgins, & Rothstein, 2009, 2010; Harwell, Maeda, Bishop, & Xie, 2006) and the current studies were randomly sampled from a larger population, allowing for estimating the average true effect in the larger population of studies (Viechtbauer, 2010).

Given that most meta-analyses incorporate a set of studies that are not identical in terms of research design and/or the characteristics of the samples, those differences may lead to variability among the true effect sizes (Viechtbauer, 2010). One way to model this variability (or heterogeneity) is using a random-effects model, which can estimate both the average true effect and the amount of variability among the true effects (Viechtbauer, 2010). Unlike the fixed-effects model that assumes variability is due only to sampling error, the random-effects model is assumed to have sampling error and a random-effects variance, which reflects heterogeneity in effect sizes due to systematic differences between studies (Rosenthal et al., 2006). Therefore, the present study used the random-effects model to account for unexplained heterogeneity within the estimated true correlation coefficient (Rosenthal et al., 2006).

In addition, a meta-analysis can be viewed as a special case of the two-level multilevel analysis within a Hierarchical Linear Model (HLM) framework (Raudenbush & Bryk, 1985). A more general linear model (i.e., mixed-effects model) includes moderators in the two-level model, also called a conditional model, which enables the researcher to account for part of the variability in the true effects (Viechtbauer, 2010). By doing so, the researcher can examine which moderators and to what extent the moderators included in the model may influence the average true effect and the amount of residual variability among the true effects that is not explained by the moderators (Maeda & Harwell, 2016; Viechtbauer, 2010). Thus, with the conditional multilevel meta-analysis, it is possible to examine which moderators (i.e., factors) influence the strength of the relation between CBM and state achievement tests of reading comprehension.

Data Analysis

First, an unconditional two-level meta-analysis was employed to yield the overall mean correlation coefficient and to examine whether variances between studies (Level 2) were statistically significant. If variances were not statistically different from zero, any additional analyses (i.e., conditional meta-analysis) would not be necessary. Second, if variances were significantly different from zero in the unconditional model, a conditional meta-analysis would be conducted. Based on the identified factors that may influence the relation between CBM and state achievement tests, traditional moderator analyses (one moderator at a time analysis) were conducted. Third, meta-regression analyses were conducted to examine what amount of variance was accounted for by the combined set of all moderators (DeCoster, 2004). All models described above were fitted using the Comprehensive Meta-Analysis (CMA) version 3.0 software (Borenstein, Hedges, Higgins, & Rothstein, 2009).

For conducting the multilevel meta-analysis, correlation coefficients from each individual study were transformed into *z* scores using Fisher's *r*-to-*z* transformation because the sampling distribution of correlation coefficients tends to be somewhat skewed (Beretvas & Pastor, 2003; Yeo, 2009). After completing all analyses, all coefficients were converted back to the original correlations for better interpretation of effect size (Yeo, 2009).

Final Studies and Correlations Included

At first, a total of 61 studies with 237 correlations were identified. However, 105 correlations were excluded because they were dependent with each other. For example, if the same participants in a study were administered CBM multiple times (e.g., fall, winter, spring), only one correlation (in spring) was included in the final analyses to avoid the issue of dependency of correlations (Harwell & Maeda, 2008). Similarly, when the same participants were administered CBM longitudinally across grades (e.g., Grade 1, Grade 2, and Grade 3), only the correlation Grade 3 was included. The rationale for these decisions was based on choosing the most concurrent (i.e., closest) relation between CBM and state tests of reading comprehension, given that the closest correlations in time likely reflects the most accurate relation.

Regarding outlier, one potential outlier was identified (Kim et al., 2010 r = .1); however, the study was included in the final analyses because the overall correlations and I^2 (between study heterogeneity index) were very similar with and without the outlier.

CHAPTER IV

RESULTS

This chapter summarizes the results of the meta-analysis on the relations between two CBM tasks (oral reading and maze) and state tests of reading comprehension. Specific research questions were as follows: (1) What are the estimated correlations between CBM and reading comprehension on state tests, and do they vary by type of task (oral reading versus maze) and grade level? (2) Are the average estimated correlations significantly different from zero? If so, to what extent are the relations between CBM and state tests of reading comprehension influenced by potential moderating variables?

In the first section, I summarize features of studies included in the meta-analysis, including: (1) characteristics of studies, (2) characteristics of participants, (3) characteristics of CBM, and (4) characteristics of state tests of reading comprehension. In the second section, I present the meta-analysis results, including: (1) the unconditional meta-analysis, (2) the conditional meta-analysis by each categorical moderator (i.e., traditional moderator analyses), and (3) a meta-regression incorporating all categorical and continuous moderators simultaneously into the model.

Summary of Studies

Sixty-one studies with 132 correlations were identified for the current meta-analysis. Below, characteristics of studies identified, participants, CBM, and state achievement tests are summarized. Also, a detailed summary of all included studies can be found in Appendix A and B.

Characteristics of Studies

The characteristics of the 61 included studies are summarized in Table 2. About 45% of the studies were published after 2010, indicating many studies on the relation between CBM to state achievement tests continued to be published within the last decade. More than half of the studies (62.3%) were journal articles.

Table 2

Summary of Study Characteristics

Category	Description of category	Number of studies (%)
Study characteristic		
Year of publication	Before 2010	34 (55.7)
	2010-2016	27 (44.3)
Type of publication	Peer-reviewed journal	38 (62.3)
	Dissertation	15 (24.6)
	Technical report	8 (13.1)

Characteristics of Participants

Participants characteristics are summarized in Table 3. About 80% of the 132 correlations were from studies conducted with elementary students. Given that a few studies had a very large sample size (e.g., *n* = 156,179 in Kim et al., 2015; *n* = 16,539 in Roehrig et al., 2008), the sample size mean was much larger than the median. The mean proportions of ELL and special education students were 13% and 14%, respectively. The mean proportion of students receiving free or reduced lunch was 49.5%.

In addition to reporting mean proportions of students for each demographic variable, studies were categorized based on whether the proportion of students fitting a specific characteristic was low (below .50 *SD* of the mean proportion), medium (within .50 *SD* of the mean proportion), or high (above .50 *SD* of the mean proportion).

Table 3

Summary of Participants Characteristics

Category	Description of category	Number of correlations (%)
Grade range	Grades 1-3 (Primary)	59 (44.7)
(*n* =132, *M* = 4.4)	Grade 4-6 (Intermediate)	46 (34.8)
	Grade 7-10 (Secondary)	27 (20.5)
Sample size	Low: less than 100	42 (31.8)
(*M* = 9255.8	Medium: between 100 and 500	42 (31.8)
Mdn = 185.5)	High: > 500	48 (36.4)
Female	Low: <46%	8 (10)
(*n* = 80, *M* = 49%,	Medium: 46% - 51%	58 (72.5)
SD = .046)	High: > 51%	14 (17.5)
ELL	Low: < 1.9%	23 (28)
(*n* = 82, *M* = 13%,	Medium: 1.9% - 24.1%	45 (54.9)
SD = .222)	High: <24.1%	14 (17.1)
Special education	Low: < 6.7%	13 (14.9)
(*n* = 87, *M* = 14%,	Medium: 6.7% - 21.3%	69 (79.3)
SD = .146)	High: > 21.3%	5 (5.8)
FRL	Low: < 38.3%	32 (30.2)
(*n* = 106, *M* = 49.5%,	Medium: 38.3% - 60.7%	40 (37.7)
SD = .224)	High: > 60.7%	34 (32.1)
White	Low: < 47.1%	38 (33.9)
(*n* = 112, *M* = 59.4%,	Medium: 47.1% - 71.6%	35 (31.3)
SD = .245)	High: >71.6%	39 (34.8)

Note. ELL = English language learners, FRL = free or reduced lunch.

Characteristics of CBM

The characteristics of CBM are summarized in Table 4. More than half of the studies (52.3%) used three CBM passages and used the mean score. Regarding the type of CBM task, the majority of studies (about 80%) examined the relation between oral reading and state tests. In terms of type of CBM passage, the majority of studies (about 84%) used standardized commercial passages (e.g., DIBELS, *easy*CBM). For type of CBM administrator, in most studies (73.7%), school personnel, such as teachers or school psychologists, administered CBM.

Table 4

Summary of CBM Characteristics

Category	Description of category	Number of correlations (%)
CBM characteristic		
Number of CBM passages used (*n* = 132)	1 Passage	51 (38.6)
	2 Passages	12 (9.1)
	3 Passages	69 (52.3)
Type of CBM task (*n* = 132)	Oral reading	103 (78)
	Maze	29 (22)
Type of CBM passage (*n* = 132)	Standardized	111 (84.1)
	Researcher developed	21 (15.9)
Who administered CBM (*n* = 114)	Researcher (e.g., PI, graduate students, research assistant)	30 (26.3)
	School personnel (e.g., teacher, school psychologist)	84 (73.7)

Characteristics of State Tests of Reading Comprehension

Regarding the development type of state tests (Table 5), only about 16% of the studies used commercially prepared achievement tests (e.g., ITBS, SAT-10), which have been used to assess student achievement in some states (Wanzek et al., 2010). Most studies (75.8%) used criterion-referenced tests. Multiple choice and mixed formats (both multiple choice and open-ended questions) were used equally across studies. The mean time interval between CBM and state tests administration was 3.6 months; the majority of studies (about 85%) administered state tests within 6 months after CBM administration. In this meta-analysis, a total of 24 different state achievement tests in 61 studies were included (see Appendix C). In addition, most studies reported the internal consistency (i.e., Cronbach's alpha) as the test reliability and it ranged *r*s = .82 to .95. However, about 25% of the studies did not provide the reliability information of the state tests.

Table 5

Summary of State Tests Characteristics

Category	Description of category	Number of correlations (%)
State achievement test characteristic		
Type of tests	Commercial	22 (16.7%)
(*n* = 132)	State-developed	110 (83.3%)
Norm- or criterion-referenced	Norm-referenced	28 (21.2%)
(*n* = 132)	Criterion-referenced	100 (75.8%)
	Mixed	4 (3%)
Response formats	Multiple choice	65 (49.2%)
(*n* = 132)	Mixed (multiple choice & open ended)	67 (50.8%)
Time interval between CBM	Within 1 month	62 (48.1%)
and state test	Between 1 and 6 months	48 (37.2%)
(*n* = 129, *M* = 3.6 months)	Over 6 months	19 (14.7%)

Meta-Analysis Results

In this section, unconditional and conditional meta-analysis results are summarized. The unconditional meta-analysis reports the estimated mean correlations for all studies and separate mean correlations for oral reading and maze with no covariates or moderators included. The conditional meta-analysis consists of a series of traditional moderator analyses with categorical moderators, followed by a meta-regression, in which models involving all categorical and continuous moderators of interest are reported.

Unconditional Analysis

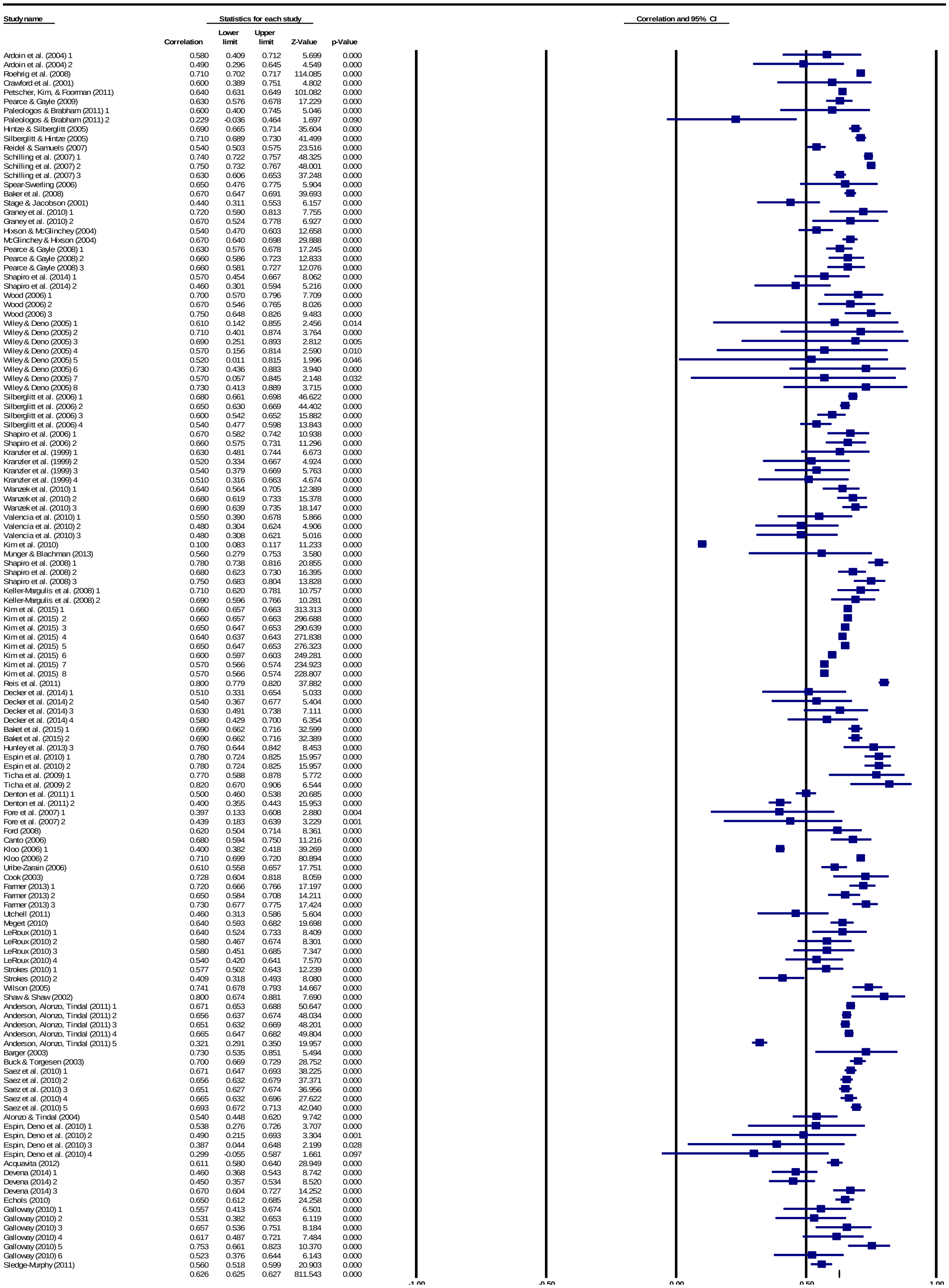
As shown in Table 6, the overall effect size for all correlations was $r = .63$ with a 95% confidence interval of $r = .62$ to .64. The *Z*-value for a test of the null hypothesis was 66.887 ($p = .000$), indicating the mean effect size (mean correlation between CBM and reading comprehension measured by state tests) was very likely to differ from zero. For oral reading, the mean correlation was $r = .63$ with a 95% confidence interval of $r = .60$ to .66. For maze, the mean correlation was $r = .60$, with a 95% confidence interval of $r = .57$ to .62. For both CBM reading tasks, the *Z*-value for a test of the null was statistically significant, which means it is likely that the mean correlation differed from zero. Estimated correlations and 95% CI for each study are also shown in the forest plot in Figure 1.

For heterogeneity, the Q-value was 12058.41 with $df = 131$ and $p < .000$ for all studies. The Q-value for oral reading and maze was 8070.96 and 3908.53, respectively, with *p*s < .000. These findings indicate that it is unlikely that all of the variance is due to sampling error; rather, the true effect size varies from study to study. The between-studies variance (T^2) was estimated as .011 for all studies, .060 for oral reading, and .004 for maze. The proportion of variance due to real differences (I^2) was 98.914 for all studies, which means that about 99% of the observed variance reflects real differences in study effects. I^2 for oral reading and maze was 98.736 and 99.283, respectively.

Table 6

Summary of Unconditional Model

	Effect size and 95% interval				Test of null	
	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
All	132	.63	.62	.64	66.887	.000
Oral reading	103	.63	.60	.66	29.243	.000
Maze	29	.60	.57	.62	40.058	.000
	Heterogeneity					
	Q-value	df(Q)	*p*-value	I^2	T^2	SE
All	12058.41	131	.000	98.91	.011	.005
Oral reading	8070.96	102	.000	98.74	.060	.019
Maze	3908.53	28	.000	99.28	.004	.002



Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Ardoin et al. (2004) 1	0.580	0.409	0.712	5.699	0.000
Ardoin et al. (2004) 2	0.490	0.296	0.645	4.549	0.000
Roehrig et al. (2008)	0.710	0.702	0.717	114.085	0.000
Crawford et al. (2001)	0.600	0.389	0.751	4.802	0.000
Petscher, Kim, & Foorman (2011)	0.640	0.631	0.649	101.082	0.000
Pearce & Gayle (2009)	0.630	0.576	0.678	17.229	0.000
Paleologos & Brabham (2011) 1	0.600	0.400	0.745	5.046	0.000
Paleologos & Brabham (2011) 2	0.229	-0.036	0.464	1.697	0.090
Hintze & Silberglitt (2005)	0.690	0.665	0.714	35.604	0.000
Silberglitt & Hintze (2005)	0.710	0.689	0.730	41.499	0.000
Reidel & Samuels (2007)	0.540	0.503	0.575	23.516	0.000
Schilling et al. (2007) 1	0.740	0.722	0.757	48.325	0.000
Schilling et al. (2007) 2	0.750	0.732	0.767	48.001	0.000
Schilling et al. (2007) 3	0.630	0.606	0.653	37.248	0.000
Spear-Swerling (2006)	0.650	0.476	0.775	5.904	0.000
Baker et al. (2008)	0.670	0.647	0.691	39.693	0.000
Stage & Jacobson (2001)	0.440	0.311	0.553	6.157	0.000
Graney et al. (2010) 1	0.720	0.590	0.813	7.755	0.000
Graney et al. (2010) 2	0.670	0.524	0.778	6.927	0.000
Hixson & McGlinchey (2004)	0.540	0.470	0.603	12.658	0.000
McGlinchey & Hixson (2004)	0.670	0.640	0.698	29.888	0.000
Pearce & Gayle (2008) 1	0.630	0.576	0.678	17.245	0.000
Pearce & Gayle (2008) 2	0.660	0.586	0.723	12.833	0.000
Pearce & Gayle (2008) 3	0.660	0.581	0.727	12.076	0.000
Shapiro et al. (2014) 1	0.570	0.454	0.667	8.062	0.000
Shapiro et al. (2014) 2	0.460	0.301	0.594	5.216	0.000
Wood (2006) 1	0.700	0.570	0.796	7.709	0.000
Wood (2006) 2	0.670	0.546	0.765	8.026	0.000
Wood (2006) 3	0.750	0.648	0.826	9.483	0.000
Wiley & Deno (2005) 1	0.610	0.142	0.855	2.456	0.014
Wiley & Deno (2005) 2	0.710	0.401	0.874	3.764	0.000
Wiley & Deno (2005) 3	0.690	0.251	0.893	2.812	0.005
Wiley & Deno (2005) 4	0.570	0.156	0.814	2.590	0.010
Wiley & Deno (2005) 5	0.520	0.011	0.815	1.996	0.046
Wiley & Deno (2005) 6	0.730	0.436	0.883	3.940	0.000
Wiley & Deno (2005) 7	0.570	0.057	0.845	2.148	0.032
Wiley & Deno (2005) 8	0.730	0.413	0.889	3.715	0.000
Silberglitt et al. (2006) 1	0.680	0.661	0.698	46.622	0.000
Silberglitt et al. (2006) 2	0.650	0.630	0.669	44.402	0.000
Silberglitt et al. (2006) 3	0.600	0.542	0.652	15.882	0.000
Silberglitt et al. (2006) 4	0.540	0.477	0.598	13.843	0.000
Shapiro et al. (2006) 1	0.670	0.582	0.742	10.938	0.000
Shapiro et al. (2006) 2	0.660	0.575	0.731	11.296	0.000
Kranzler et al. (1999) 1	0.630	0.481	0.744	6.673	0.000
Kranzler et al. (1999) 2	0.520	0.334	0.667	4.924	0.000
Kranzler et al. (1999) 3	0.540	0.379	0.669	5.763	0.000
Kranzler et al. (1999) 4	0.510	0.316	0.663	4.674	0.000
Wanzek et al. (2010) 1	0.640	0.564	0.705	12.389	0.000
Wanzek et al. (2010) 2	0.680	0.619	0.733	15.378	0.000
Wanzek et al. (2010) 3	0.690	0.639	0.735	18.147	0.000
Valencia et al. (2010) 1	0.550	0.390	0.678	5.866	0.000
Valencia et al. (2010) 2	0.480	0.304	0.624	4.906	0.000
Valencia et al. (2010) 3	0.480	0.308	0.621	5.016	0.000
Kim et al. (2010)	0.100	0.083	0.117	11.233	0.000
Munger & Blachman (2013)	0.560	0.279	0.753	3.580	0.000
Shapiro et al. (2008) 1	0.780	0.738	0.816	20.855	0.000
Shapiro et al. (2008) 2	0.680	0.623	0.730	16.395	0.000
Shapiro et al. (2008) 3	0.750	0.683	0.804	13.828	0.000
Keller-Margulis et al. (2008) 1	0.710	0.620	0.781	10.757	0.000
Keller-Margulis et al. (2008) 2	0.690	0.596	0.766	10.281	0.000
Kim et al. (2015) 1	0.660	0.657	0.663	313.313	0.000
Kim et al. (2015) 2	0.660	0.657	0.663	296.688	0.000
Kim et al. (2015) 3	0.650	0.647	0.653	290.639	0.000
Kim et al. (2015) 4	0.640	0.637	0.643	271.838	0.000
Kim et al. (2015) 5	0.650	0.647	0.653	276.323	0.000
Kim et al. (2015) 6	0.600	0.597	0.603	249.281	0.000
Kim et al. (2015) 7	0.570	0.566	0.574	234.923	0.000
Kim et al. (2015) 8	0.570	0.566	0.574	228.807	0.000
Reis et al. (2011)	0.800	0.779	0.820	37.882	0.000
Decker et al. (2014) 1	0.510	0.331	0.654	5.033	0.000
Decker et al. (2014) 2	0.540	0.367	0.677	5.404	0.000
Decker et al. (2014) 3	0.630	0.491	0.738	7.111	0.000
Decker et al. (2014) 4	0.580	0.429	0.700	6.354	0.000
Baket et al. (2015) 1	0.690	0.662	0.716	32.599	0.000
Baket et al. (2015) 2	0.690	0.662	0.716	32.389	0.000
Hunley et al. (2013) 3	0.760	0.644	0.842	8.453	0.000
Espin et al. (2010) 1	0.780	0.724	0.825	15.957	0.000
Espin et al. (2010) 2	0.780	0.724	0.825	15.957	0.000
Ticha et al. (2009) 1	0.770	0.588	0.878	5.772	0.000
Ticha et al. (2009) 2	0.820	0.670	0.906	6.544	0.000
Denton et al. (2011) 1	0.500	0.460	0.538	20.685	0.000
Denton et al. (2011) 2	0.400	0.355	0.443	15.953	0.000
Fore et al. (2007) 1	0.397	0.133	0.608	2.880	0.004
Fore et al. (2007) 2	0.439	0.183	0.639	3.229	0.001
Ford (2008)	0.620	0.504	0.714	8.361	0.000
Canto (2006)	0.680	0.594	0.750	11.216	0.000
Kloo (2006) 1	0.400	0.382	0.418	39.269	0.000
Kloo (2006) 2	0.710	0.699	0.720	80.894	0.000
Uribe-Zarain (2006)	0.610	0.558	0.657	17.751	0.000
Cook (2003)	0.728	0.604	0.818	8.059	0.000
Farmer (2013) 1	0.720	0.666	0.766	17.197	0.000
Farmer (2013) 2	0.650	0.584	0.708	14.211	0.000
Farmer (2013) 3	0.730	0.677	0.775	17.424	0.000
Utchell (2011)	0.460	0.313	0.586	5.604	0.000
Megert (2010)	0.640	0.593	0.682	19.698	0.000
LeRoux (2010) 1	0.640	0.524	0.733	8.409	0.000
LeRoux (2010) 2	0.580	0.467	0.674	8.301	0.000
LeRoux (2010) 3	0.580	0.451	0.685	7.347	0.000
LeRoux (2010) 4	0.540	0.420	0.641	7.570	0.000
Strokes (2010) 1	0.577	0.502	0.643	12.239	0.000
Strokes (2010) 2	0.409	0.318	0.493	8.080	0.000
Wilson (2005)	0.741	0.678	0.793	14.667	0.000
Shaw & Shaw (2002)	0.800	0.674	0.881	7.690	0.000
Anderson, Alonzo, Tindal (2011) 1	0.671	0.653	0.688	50.647	0.000
Anderson, Alonzo, Tindal (2011) 2	0.656	0.637	0.674	48.034	0.000
Anderson, Alonzo, Tindal (2011) 3	0.651	0.632	0.669	48.201	0.000
Anderson, Alonzo, Tindal (2011) 4	0.665	0.647	0.682	49.804	0.000
Anderson, Alonzo, Tindal (2011) 5	0.321	0.291	0.350	19.957	0.000
Barger (2003)	0.730	0.535	0.851	5.494	0.000
Buck & Torgesen (2003)	0.700	0.669	0.729	28.752	0.000
Saez et al. (2010) 1	0.671	0.647	0.693	38.225	0.000
Saez et al. (2010) 2	0.656	0.632	0.679	37.371	0.000
Saez et al. (2010) 3	0.651	0.627	0.674	36.956	0.000
Saez et al. (2010) 4	0.665	0.632	0.696	27.622	0.000
Saez et al. (2010) 5	0.693	0.672	0.713	42.040	0.000
Alonzo & Tindal (2004)	0.540	0.448	0.620	9.742	0.000
Espin, Deno et al. (2010) 1	0.538	0.276	0.726	3.707	0.000
Espin, Deno et al. (2010) 2	0.490	0.215	0.693	3.304	0.001
Espin, Deno et al. (2010) 3	0.387	0.044	0.648	2.199	0.028
Espin, Deno et al. (2010) 4	0.299	-0.055	0.587	1.661	0.097
Acquavita (2012)	0.611	0.580	0.640	28.949	0.000
Devena (2014) 1	0.460	0.368	0.543	8.742	0.000
Devena (2014) 2	0.450	0.357	0.534	8.520	0.000
Devena (2014) 3	0.670	0.604	0.727	14.252	0.000
Echols (2010)	0.650	0.612	0.685	24.258	0.000
Galloway (2010) 1	0.557	0.413	0.674	6.501	0.000
Galloway (2010) 2	0.531	0.382	0.653	6.119	0.000
Galloway (2010) 3	0.657	0.536	0.751	8.184	0.000
Galloway (2010) 4	0.617	0.487	0.721	7.484	0.000
Galloway (2010) 5	0.753	0.661	0.823	10.370	0.000
Galloway (2010) 6	0.523	0.376	0.644	6.143	0.000
Sledge-Murphy (2011)	0.560	0.518	0.599	20.903	0.000
	0.626	0.625	0.627	811.543	0.000

Figure 1. Forest Plot for the Estimated Correlations

Conditional Analysis (one moderator at a time)

Because the unconditional analysis revealed that the true effect size (i.e., correlation) varied from study to study, conditional analyses were conducted to examine the effects of potential moderators that could explain the observed variance. As a preliminary analysis, a series of traditional one-moderator-at-a-time analyses were done to analyze the effects of each of the following categorical covariates (i.e., moderators): type of publication, grade level, grade range (primary, intermediate, secondary), student demographics (percent female, ELL, special education, FRL, White), type of CBM, type of CBM passage, number of CBM passages, type of CBM administrator, development type of state tests (commercial or state-developed), test type of state test (norm- or criterion-referenced), and response format of state tests. For student demographics, percentages were transformed into three categories (low, medium, high) based on the criteria described in “Characteristics of Participants.”

The effects of type of publication are shown in Table 7. The mean correlation was r = .64 for journal articles, r = .60 for dissertations, and r = .64 for tech reports; all mean correlations were statistically significant (ps = .000). The mean correlation was the same for journal articles and tech reports, and lowest for dissertations. The Q-value for this difference in correlations was 4.012 (df = 2, p-value = .134), indicating that there is no evidence that the mean correlations reliably differed by publication type. The variance of true correlations (T^2) across subgroups (i.e., publication types combined) was .011. In addition, the combined estimate of I^2--the proportion of the variance in observed effects that is due to variation in true effects (Borenstein, Higgins, Hedges, & Rothstein, 2009)--

was 98.92, indicating that most of the within-subgroup variance reflects real differences in study effects. Further, to test the assumption that all studies within a subgroup share a common correlation, Q-values were summed up across subgroups; the overall Q-value was 11946.28 with *p* = .000. This Q-value indicates that the true correlations vary from study to study within subgroups. In other words, knowing whether a study is of a specific publication type does not completely predict its correlation.

Table 7

Effects of Type of Publication

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Journal article	84	.64	.62	.65	53.060	.000
Dissertation	29	.60	.57	.63	30.508	.000
Tech report	19	.64	.60	.67	26.791	.000
Overall	132	.63	.62	.64	66.782	.000

Heterogeneity					
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2
4.012	2	.134	11946.28 (*p* = .00)	98.92	.011

Mean correlations for each grade level are summarized in Table 8 (oral reading and maze together), Table 9 (oral reading only), and Table 10 (maze only).

Table 8

Estimated Effect Sizes for Each Grade Level: Oral Reading and Maze

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Grade 1	7	.59	.48	.69	3.242	.000
Grade 2	8	.62	.55	.72	10.259	.000
Grade 3	44	.63	.59	.67	22.328	.000
Grade 4	18	.65	.59	.70	15.471	.000
Grade 5	19	.63	.56	.69	13.921	.000
Grade 6	9	.61	.51	.68	10.205	.000
Grade 7	13	.56	.47	.63	10.655	.000
Grade 8	12	.67	.60	.73	12.769	.000
Grade 9	1	.57	.25	.78	3.242	.001
Grade 10	1	.57	.25	.78	3.242	.001
Overall	132	.62	.59	.65	27.592	.000

Table 9

Estimated Effect Sizes for Each Grade Level: Oral Reading

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Grade 1	7	.59	.44	.71	6.424	.000
Grade 2	8	.64	.51	.74	7.756	.000
Grade 3	38	.64	.58	.69	16.324	.000
Grade 4	15	.65	.56	.72	10.751	.000
Grade 5	14	.64	.55	.72	9.793	.000
Grade 6	7	.62	.48	.73	7.012	.000
Grade 7	8	.58	.43	.69	6.659	.000
Grade 8	6	.69	.55	.79	7.206	.000
Overall	103	.63	.60	.67	24.099	.000

Table 10

Estimated Effect Sizes for Each Grade Level: Maze

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Grade 3	6	.58	.44	.69	6.668	.000
Grade 4	3	.65	.49	.77	6.376	.000
Grade 5	5	.57	.41	.70	5.798	.000
Grade 6	2	.54	.32	.70	4.323	.000
Grade 7	5	.53	.38	.64	6.264	.000
Grade 8	6	.66	.55	.74	8.962	.000
Grade 9	1	.57	.26	.77	3.346	.000
Grade 10	1	.57	.26	.77	3.346	.000
Overall	29	.59	.52	.65	13.724	.000

Thus far, I have provided descriptive results for mean correlations for oral reading, maze, and both tasks for each grade. However, due to the very small number of correlations for some grades (e.g., grades 9 and 10), which may influence the precision of the estimate (Borenstein et al., 2009), analyses of the effects of grade level were conducted using grade *range*. That is, the following analyses examined whether the correlations between CBM and reading comprehension on state tests differed by primary, intermediate, and secondary grade levels. Previous researchers have suggested that the oral reading task is a better indicator of reading comprehension on criterion measures for primary grades (Grades 1-3), and that maze might be better for intermediate (Grades 4-6) and secondary grades (Wayman et al., 2007). Thus, analyzing the effect of grade range will provide more empirical evidence about these relations.

First, the effects of grade range for both types of CBM (oral reading and maze together) are shown in Table 11. The mean correlation was *r* = .63 for primary grades (Grades 1-3), *r* = .65 for intermediate grades (Grades 4-6), and *r* = .60 for secondary (middle and high school); all mean correlations differed significantly from zero (*p*s = .000). The mean correlation was highest for intermediate grades and lowest for secondary grades. This difference was statistically significant (Q = 8.244, *df* = 9, *p* = .016), indicating that the mean correlation differed by the grade range. The summed Q-value across subgroups was 10049.34 (*p* = .00), indicating that the true correlations vary from study to study within subgroups (primary, intermediate, and secondary). In other words, knowing whether a study falls into each grade range does not completely predict its correlation.

Table 11

Effects of Grade Range: Both Oral Reading and Maze

Effect size and 95% interval					Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Primary (Grades 1-3)	59	.63	.61	.65	43.971	.000
Intermediate (Grades 4-6)	43	.65	.63	.67	38.752	.000
Secondary (Grades 7-10)	30	.60	.57	.62	29.472	.000
Overall	132	.63	.60	.66	29.751	.000

Heterogeneity					
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2
8.244	2	.016	10049.34 (*p* = .00)	98.72	.011

Next, the effects of grade range for oral reading only was examined and the results are shown in Table 12. The mean correlation was *r* = .63 for primary grades (Grades 1-3), *r* = .65 for intermediate grades (Grades 4-6), and *r* = .61 for secondary (middle and high school); all mean correlations differed significantly from zero (*p*s = .000). The mean correlation was the highest for intermediate grades and lowest for secondary grades. This difference was not statistically significant (Q = .513, *df* = 2, *p* = .773), indicating that there is no evidence that the mean correlations reliably differed by grade range. The summed Q-value across subgroups was 7790.65 (*p* = .00), indicating that the true correlations for oral reading vary across studies.

Table 12

Effects of Grade Range: Oral Reading

	Effect size and 95% interval				Test of null	
Category	Number of correlation	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Primary (Grades 1-3)	53	.63	.59	.67	20.691	.000
Intermediate (Grades 4-6)	34	.65	.59	.70	17.158	.000
Secondary (Grades 7-10)	16	.61	.53	.69	10.959	.000
Overall	103	.64	.60	.66	29.019	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
.513	2	.773	7790.65 (*p* = .00)		98.72	.011

The effects of grade range for maze only were also examined and results are shown in Table 13. The mean correlation was *r* = .62 for primary grades (Grades 1-3), *r* = .64 for intermediate grades (Grades 4-6), and *r* = .57 for secondary (middle and high school); all mean correlations were statistically significant (*p*s = .000). Similar to the results for oral reading, the mean correlation was highest for the intermediate grades and lowest for secondary grades. In contrast to oral reading results, however, this difference was statistically significant (Q = 11.34, *df* = 2, *p* = .003), indicating that the mean correlation varied by grade range. The summed Q-value across subgroups was 1660.47 with *p*-value = .00, indicating that the true correlations for maze vary across studies.

Table 13

Effects of Grade Range: Maze

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Primary (Grades 1-3)	6	.62	.57	.67	17.945	.000
Intermediate (Grades 4-6)	9	.64	.61	.67	29.217	.000
Secondary (Grades 7-10)	14	.57	.55	.60	35.158	.000
Overall	29	.61	.56	.66	18.288	.000
			Heterogeneity			
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2	
11.340	2	.003	1660.47 (*p* = .00)	98.72	.011	

Next, the effects of five student demographic variables were examined (Table 14). For each demographic variable, the proportion in the sample was categorized as high (above .50 *SD* of the mean proportion), medium (within .50 *SD* of the mean proportion), or low (below .50 *SD* of the mean proportion), and correlations for each level were estimated.

Regarding the proportion of females, the mean correlation was highest for studies with medium proportions of females (i.e., studies with relatively similar proportions of males and females; *r* = .63), and lowest for the low proportion (i.e., studies with smaller proportions of females; *r* = .60). This difference, however, was not statistically significant (Q = 1.647, *df* = 2, *p* = .65), indicating that there is no evidence that the mean correlations reliably differed by the proportion of female students. The summed Q-value across subgroups was 11973.8 with *p*-value = .00, indicating that the true correlations varied across studies, even within each subgroup.

For the proportion of ELLs, the mean correlation was highest for the medium proportion (*r* = .65), and lowest for the low proportion (*r* = .61). This difference was not statistically significant (Q = 7.145, *df* = 3, *p* = .07), indicating that there is no evidence that the mean correlations reliably differed by the proportion of ELL students.

For the proportion of special education status, the mean correlation was highest for the medium proportion (*r* = .65), and lowest for the high proportion (*r* = .52). This difference was statistically significant (Q = 25.289, *df* = 3, *p* = .00), indicating that the mean correlation differed by the proportion of special education students.

Regarding the proportion of FRL, the mean correlation was highest for the medium proportion (*r* = .65), and lowest for the low proportion (*r* = .60). This difference was statistically significant (Q = 14.195, *df* = 3, *p* = .00), indicating that the mean correlation differed by the proportion of students receiving FRL.

For the proportion of White students, the mean correlation was highest for the high proportion (*r* = .64), and lowest for the medium proportion (*r* = .62). This difference, however, was not statistically significant (Q = 1.141, *df* = 3, *p* = .77), indicating that there is no evidence that the mean correlations reliably differed by the proportion of White students.

Table 14

Effects of Student Demographic Information

		Effect size and 95% interval				Test of null	
Category		Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Female % (*n* = 80)	Low	8	.60	.54	.66	14.329	.000
	Medium	58	.63	.61	.65	47.476	.000
	High	14	.62	.57	.66	19.386	.000
ELL % (*n* = 82)	Low	23	.61	.58	.64	26.401	.000
	Medium	45	.65	.63	.66	46.991	.000
	High	14	.65	.60	.69	20.968	.000
SpEd % (*n* = 87)	Low	13	.56	.51	.60	18.895	.000
	Medium	69	.65	.63	.66	57.774	.000
	High	5	.52	.42	.61	8.533	.000
FRL % (*n* = 106)	Low	32	.60	.58	.63	34.091	.000
	Medium	40	.65	.63	.67	48.299	.000
	High	34	.64	.62	.66	28.337	.000
White % (*n* = 112)	Low	38	.63	.61	.65	39.871	.000
	Medium	35	.62	.59	.62	33.822	.000
	High	39	.64	.61	.67	20.756	.000

Heterogeneity						
	Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2
Female %	1.647	3	.65	11973.8 (*p* = .00)	99.25	.01
ELL %	7.145	3	.07	10432.08 (*p* = .00)	98.89	.006
SpEd %	25.289	3	.00	10154.99 (*p* = .00)	98.73	.006
FRL %	14.195	3	.00	7298.71 (*p* = .00)	98.57	.007
White %	1.141	3	.77	11893.92 (*p* = .00)	99.00	.01

Note. *n* = the total number of correlations.

Next, the results for the effects of type of CBM (oral reading and maze) are shown in Table 15. The mean correlation was *r* = .64 for oral reading and *r* = .59 for maze, and this difference was statistically significant (Q = 6.961, *df* = 1, *p* = .008), indicating that the mean correlation differed by the type of CBM. The summed Q-value across subgroups was 11,979.50 (*p* = .00), indicating that the true correlations varied across studies by type of CBM. Put another way, whether a study included oral reading or maze does not completely predict its correlation.

Table 15

Effects of Type of CBM

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Oral reading	103	.64	.62	.65	60.733	.000
Maze	29	.59	.56	.62	27.274	.000
Overall	132	.62	.57	.66	19.67	.000
			Heterogeneity			
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2	
6.961	1	.008	11979.5 (*p* = .00)	98.91	.011	

In addition, to examine whether the effects of type of CBM differ between primary, intermediate, and secondary grade ranges, three separate moderator analyses were conducted. As shown in Tables 16, 17, and 18, for primary grades (Grades 1-3), estimated mean correlations between oral reading and state tests of reading comprehension (*r* = .63) were higher than those for maze (*r* = .58). However, there was no significant effect of type of CBM for primary grades (*p* = .525). For intermediate grades (Grades 4-6), correlations were *r* = .66 for oral reading and *r* = .65 for maze. For secondary grades (Grades 7-10), estimated mean correlations were *r* = .61 for oral reading and *r* = .59 for maze. Again, type of CBM was not significant moderator for intermediate (*p* = .10) and secondary grades (*p* = .11), respectively. Finally, one additional moderator analysis was conducted to examine the effect of type of CBM for intermediate and secondary grades combined to address the second hypothesis (i.e., estimated average correlation between maze and reading comprehension will be larger than that between oral reading and reading comprehension for intermediate and secondary grades). As shown in Table 19, the estimated mean correlation for oral reading (*r* = .64) was larger than that for maze (*r* = .59), and this difference was significant (*p* = .001).

Table 16

Effects of Type of CBM for Primary Grades

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Oral reading	53	.63	.58	.68	18.727	.000
Maze	6	.58	.38	.72	5.074	.000
Overall	59	.63	.58	.67	19.392	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
0.404	1	.525	6942.103 (*p* = .00)		99.23	.047

Table 17

Effects of Type of CBM for Intermediate Grades

Effect size and 95% interval					Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Oral reading	34	.66	.65	.67	82.317	.000
Maze	9	.65	.63	.66	46.582	.000
Overall	43	.66	.64	.67	50.745	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
2.711	1	.10	304.060 (*p* = .00)		86.69	.001

Table 18

Effects of Type of CBM for Secondary Grades

Effect size and 95% interval					Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Oral reading	16	.61	.58	.64	28.72	.000
Maze	14	.57	.54	.61	26.03	.000
Overall	30	.59	.56	.63	24.29	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
2.537	1	.11	2204.958 (*p* = .00)		98.69	.005

Table 19

Effects of Type of CBM for Intermediate and Secondary Grades (Combined)

Effect size and 95% interval					Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Oral reading	50	.64	.63	.66	55.95	.000
Maze	23	.59	.57	.62	33.93	.000
Overall	73	.62	.57	.67	18.06	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
10.90	1	.001	4393.62 (*p* = .00)		98.37	.005

Next, the effects of type of CBM passages (standardized or researcher-developed) were examined and the results are shown in Table 20. The mean correlation was *r* = .63 for standardized passages and *r* = .64 for researcher-developed passages, but this difference was not statistically significant (Q = .734, *df* = 1, *p* = .39). The summed Q-value across subgroups was 11980.09 (*p* = .00), indicating that the true correlations varied across studies within each subgroup (studies with standardized or researcher-developed passages).

Table 20

Effects of Type of CBM Passages

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Standardized	111	.63	.61	.64	62.048	.000
Researcher-developed	21	.64	.61	.68	25.342	.000
Overall	132	.63	.62	.64	67.018	.000
			Heterogeneity			
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
.734	1	.39	11980.09 (*p* = .00)		98.91	.011

The effects of the number of CBM passages (1, 2, or 3) were examined and the results are shown in Table 21. The mean correlation was *r* = .62 for 1 passage, *r* = .61 for 2 passages, and *r* = .64 for 3 passages. Differences were not statistically significant (Q = 1.622, *df* = 2, *p* = .44). The summed Q-value across subgroups was 11526.01 (*p* = .00), indicating that the true correlations varied across studies within each subgroup (1, 2, or 3 passaged used).

Table 21

Effects of Number of CBM Passages

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
1 passage	51	.62	.60	.64	42.027	.000
2 passages	12	.61	.57	.65	21.778	.000
3 passages	69	.64	.62	.65	49.639	.000
Overall	132	.63	.61	.64	56.763	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
1.622	2	.444	11313.63 (*p* = .00)		98.86	.010

The effects of type of CBM administrator (researchers, school personnel) were examined and the results are shown in Table 22. The mean correlation was *r* = .63 for researchers and *r* = .64 for school personnel, but this difference was not statistically significant (Q = 3.516, *df* = 1, *p* = .172). The summed Q-value across subgroups was 11526.01 with *p*-value = .00, indicating that the true correlations vary from study to study within each subgroup. In other words, whether a study used researchers or school personnel to administer CBM does not completely predict its correlation.

Table 22

Effects of Type of CBM Administrator

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Researcher	30	.63	.60	.659	29.236	.000
School Personnel	84	.64	.62	.650	55.659	.000
Overall	132	.62	.60	.646	24.861	.000
Heterogeneity						

Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2
3.516	2	.172	11526.01 (*p* = .00)	98.88	.011

The effects of the development type of state tests (commercial or state-developed) were examined and the results are shown in Table 23. The mean correlation was *r* = .607 for commercially-developed tests and *r* = .633 for state-developed tests, but this difference was not statistically significant (Q = 2.119, *df* = 1, *p* = .145). The summed Q-value across subgroups was 11135.84 with *p*-value = .00, indicating that the true correlations varied across studies within each subgroup (i.e., studies using commercially- or state-developed tests).

Table 23

Effects of Development Type of State Tests

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Commercial	22	.61	.57	.64	26.377	.000
State-developed	110	.63	.62	.65	63.396	.000
Overall	132	.62	.60	.65	36.362	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2	
2.119	1	.145	11135.84 (*p* = .00)	98.83	.010	

The effects of the type of state tests (norm- or criterion-referenced) were examined and the results are shown in Table 24. The mean correlation was *r* = .62 for norm-referenced tests, *r* = .63 for criterion-referenced tests, and *r* = .58 for mixed tests, but differences were not statistically significant (Q = 2.689, *df* = 2, *p* = .261). The summed Q-value across subgroups was 11179.38 with *p*-value = .00, indicating that the true correlations varied across studies within each subgroup (i.e., studies using norm- or criterion-referenced state tests).

Table 24

Effects of Type of State Tests (Norm- or Criterion-Referenced)

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Norm-referenced	28	.62	.59	.65	30.521	.000
Criterion-referenced	100	.63	.62	.65	60.367	.000
Mixed	4	.58	.50	.65	11.024	.000
Overall	132	.62	.60	.65	37.666	.000
Heterogeneity						
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups		I^2	T^2
2.689	2	.261	11179.38 (*p* = .00)		98.85	.010

The effects of the response format of state tests (multiple choice or mixed type) were examined and the results are shown in Table 25. The mean correlation was *r* = .62 for multiple-choice type and *r* = .64 for mixed type (i.e., combination of multiple choice and open-ended questions), but this difference was marginally significant (Q = 3.754, *df* = 2, *p* = .052). Again, the summed Q-value across subgroups was 12055.34 with *p*-value = .00, indicating that the true correlations vary from study to study within each subgroup (i.e., studies using tests with multiple choice or mixed response formats).

Table 25

Effects of Response Format of State Tests

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Multiple choice	65	.62	.60	.63	46.345	.000
Mixed	67	.64	.62	.66	47.901	.000
Overall	132	.63	.60	.65	24.392	.000

Heterogeneity					
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2
3.754	1	.052	12055.34 (*p* = .00)	98.92	.011

Finally, the effects of the time interval between CBM and state test administration (within 1 month, between 1-6 months, and over 6 months) were examined and the results are shown in Table 26. The mean correlation was *r* = .65 for within 1 month, *r* = .63 for between 1 and 6 months, and *r* = .61 for over 6 months. Differences in correlations were statistically significant (Q = 16.409, *df* = 3, *p* = .001). The summed Q-value across subgroups was 9833.168 with *p*-value = .00, indicating that the true correlations vary from study to study within each subgroup and knowing whether a study include a short, medium, or long time interval between CBM and state test administration does not completely predict its correlation.

Table 26

Effects of Time Interval (categorized) between CBM and State Tests

	Effect size and 95% interval				Test of null	
Category	Number of correlations	Point estimate	Lower limit	Upper limit	*Z*-value	*p*-value
Short (within 1 month)	40	.65	.63	.67	45.367	.000
Medium (between 1-6 months)	39	.63	.61	.65	37.367	.000
Long (Over 6 months)	50	.61	.59	.63	40.644	.000
Overall	132	.62	.58	.65	25.038	.000

Heterogeneity					
Q-value	df(Q)	*p*-value	Sum of Q-values across subgroups	I^2	T^2
16.409	3	.001	9833.168 (*p* = .00)	98.72	.009

To summarize, for the traditional one-moderator-at-a-time analyses, the effects of the following moderators were examined: type of publication, grade range, student demographics (proportions of females, ELLs, students in special education, students receiving FRL, and White students), type of CBM (oral reading or maze), type of CBM passages (commercial or researcher-developed), number of CBM passages, type of CBM administrator, development type of state tests (commercial or state), type of state test (norm- or criterion-referenced), response format of state tests (multiple-choice or mixed), and time interval between CBM and state tests (short, medium, or long).

Among these variables, grade range for both oral reading and maze (*p* = .016), grade range for maze only (*p* = .003), special education proportion (*p* = .00), FRL proportion (*p* = .00), type of CBM (*p* = .008), and time interval (*p* = .001) were significant moderators. The response format of state tests was a marginally significant moderator (*p* = .052). Regarding the effect of grade range, the mean correlation was highest for intermediate grades and lowest for the secondary grades. For the effect of special education proportion, studies that included a medium range of students in special education (6.7%-21.3%) produced the highest correlation whereas studies that included high proportions (over 21.3%) produced the lowest correlation. In terms of the effect of FRL proportion, studies that included a medium range of students on FRL (38.3%-60.7%) produced the highest correlation, and low range (less than 38.3%) the lowest correlation. Regarding the effect of type of CBM, when included as the only moderator, the mean correlation between oral reading and state tests was higher than that between maze and state tests. Finally, a short time interval (within 1 month) between CBM and state test administration produced the highest correlation whereas a long time interval (over 6 months) produced the lowest correlation.

Meta-Regression

So far, each moderator's effect was investigated individually using traditional moderator analysis. Next, meta-regression was conducted to (a) examine the effects of each moderator (covariate) when other moderators were held constant and (b) determine whether the observed variance could be explained by all moderators (categorical and continuous moderators) simultaneously in the model. In this case, meta-regression is identical to multiple-regression in primary studies, except that the covariates are at the level of the study rather than at the level of the subject (Borenstein et al., 2009).

Given the considerable amount of missing values in student demographic information, whereas there were few missing values for other moderators, two meta-regression analyses were conducted. Model 1 includes all moderators except for student demographics (*n* = 129), and Model 2 added student demographics to Model 1. Due to the software's list-wise deletion for dealing with missing data, Model 2 included only 35 correlations. Although Model 2 may be considered the final model because it included every covariate, the results should be interpreted with caution because the model included a limited subset of the studies, and therefore the data do not fully represent the original data set. All estimates are reported in *z*-values for the meta-regression analyses.

Meta-regression results for Model 1 are summarized in Table 27. Among a total of nine moderators, the following five moderators were significant sources of variance in study effect sizes (correlations): type of publication (*p* = .025), grade range (*p* = .007), type of CBM (*p* = .029), development type of state test (*p* = .001), and time interval between CBM and state test administration (*p* = .000). Specifically, regarding type of publication, correlations from journal articles were significantly higher than those from dissertations (*p* = .02) whereas there was not a significant difference in correlations between journal articles and tech reports (*p* = .059). In terms of grade range differences, correlations from primary and intermediate grades were not significantly different (*p* = .954), but correlations from primary grades were significantly higher than those from secondary grades (*p* = .007). For type of CBM, when other moderators were held constant, the regression coefficient was -.079 (*p* = .029), meaning that correlations between oral reading and reading comprehension on state tests were higher than those between maze and reading comprehension. As for development type of state test, correlations from state-developed tests were significantly higher than those from commercially prepared tests of reading comprehension (.117, *p* = .001). Last, correlations were higher if the time interval between CBM and state test administration was shorter (-.009, *p* = .000), as shown in Figure 2.

Other moderators' effects (development type of CBM passage, number of CBM passages used, type of CBM administrator, and response format of state tests) were not statistically significant. That is, correlations were not significantly different whether commercially prepared or researcher-developed CBM was used (*p* = .657). In addition, correlations did not differ significantly by the number of CBM passages used (*p* = .366). Similarly, type of administrator of CBM (researchers or school personnel) did not influence the variance between correlations (*p* = .167). Further, correlations did not differ by response format (multiple choice or mixed format; *p* = .729).

Overall, R^2 (an index based on the percent reduction in true variance) for Model 1 was .28, which means that about 28% of total between-study variance was explained by the covariates (moderators) in the model. To further examine a possible interaction effect between grade range and type of CBM, an interaction term (grade range × type of CBM) was added to the Model 1. However, the interaction effect (.023, *p* = .531) was not statistically significant, which indicates that there was no significant effect of grade range depending on the type of CBM (oral reading or maze). Given that the interaction effect was not significant and to keep a parsimonious model, Model 1 was retained without the interaction term.

Table 27

Meta-Regression Results for Model 1 (All Moderators except Student Demographics, n = 129)

Covariate	Coefficient	SE	95% C.I. [low, high]	*Z*	*p*	*p* as a set
Intercept	.736	.041	[.657, .816]	18.14	.000	-
Type of publication (Journal vs. Dissertation)	-.064	.028	[-.118, -.01]	-2.32	.020	.025
Type of publication (Journal vs. Tech report)	-.064	.034	[-.131, -.002]	-1.89	.059	
Grade range (Primary vs. Intermediate)	-.001	.027	[-.054, .051]	-.06	.954	.008
Grade range (Primary vs. secondary)	-.083	.031	[-.143, -.023]	-2.71	.007	
Type of CBM (oral reading vs. maze)	-.079	.036	[-.149, -.008]	-2.18	.029	-
Development type of CBM passage (commercial vs. researcher-developed)	.015	.035	[-.053, .083]	.44	.657	-
Number of CBM passages (one vs. two)	.055	.042	[-.027, .138]	1.31	.189	.366
Number of CBM passages (one vs. three)	.019	.026	[-.031, .07]	.75	.451	
Type of CBM administrator (Researcher vs. teachers)	-.015	.03	[-.074, .044]	-.50	.619	.167
Type of CBM administrator (Researcher vs. N/A)	-.067	.039	[-.144, .009]	-1.72	.085	
Development type of state test (commercial vs. state-developed)	.117	.036	[.048, .188]	3.3	.001	-
Response format of state test (Multiple choice vs. mixed format)	.009	.026	[-.042, .061]	.35	.729	-
Time interval between CBM and state test	-.009	.002	[-.013, -.005]	-4.68	.000	-

Model Statistics					
Test of the model			Q	df	*p*-value
			61.44	13	.000
Goodness of fit	T^2	I^2(%)	Q	df	*p*-value
	.0078	98.11	6018.51	115	.000
Total between-study variance	T^2	I^2(%)	Q	df	*p*-value
	.0108	98.94	12046.10	128	.000
Proportion of total between-study variance explained by model (R^2)				.28	

Note. N/A for type of CBM administer indicates correlations with no information.

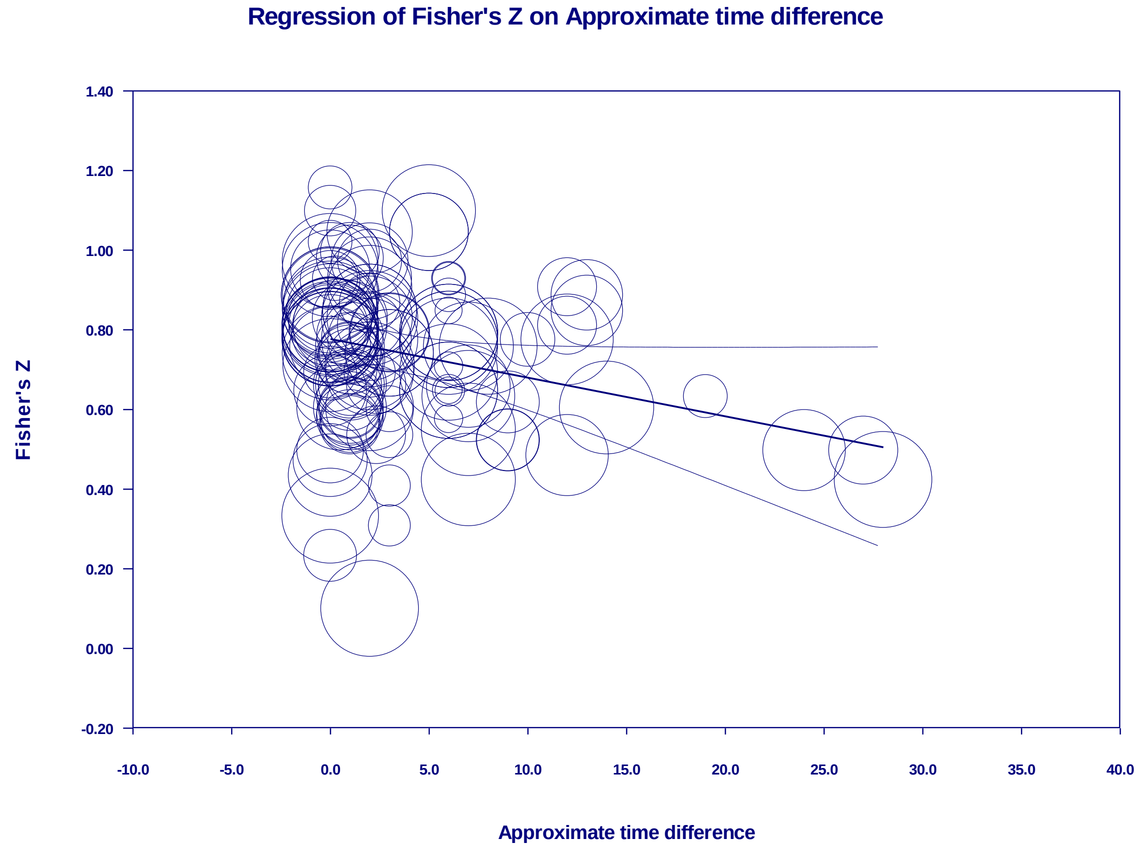


Figure 2. Scatter Plot of Regression on Time interval

The scatter plot in Figure 2 shows the relation between time interval (X-axis) and correlations in Fisher's *Z* (Y-axis) with a regression line and its confidence interval. In this plot, the size of each circle indicates relative sample size, and one could see that, in many studies, the time interval between CBM and state tests was within 5 months. This plot indicates that as time difference between CBM and state tests administration increases, the strength of the relations (correlations) decrease. Plots for other covariates are provided in Appendix D.

The Model 2 meta-regression results with moderators including student demographic information are summarized in Table 28. Compared to the former meta-regression without student demographics, Model 2 had only 35 correlations because there were many missing values for demographic information.

When student demographics were included in the model, proportion of females, special education, and FRL were significant moderators. For female percentage, the coefficient was -2.75 (*p* = .024), which means that as one unit of female percentage increased in the sample, the correlation between CBM and reading comprehension on state tests decreased by -2.75. In contrast, special education percentage in the sample showed a positive effect on the correlations (2.15, *p* = .005), indicating that studies including a higher percentage of special education students reported higher correlations. Regarding FRL percentage, studies that included a lower percentage of students on FRL produced higher correlations (-.57, *p* = .019). Other demographic variables (proportion of ELLs and White students) were not significant moderators.

As compared to Model 1, somewhat different results for the five moderators' effects were found. For type of CBM, there was no significant effect in Model 2 ($p = .229$), whereas a significant effect was found in Model 1. Regarding four other moderators (type of CBM passage, number of CBM passage, type of CBM administrator, response format of state achievement tests), significant effects were found in Model 2 (all *p*s < .05), whereas there were no significant effects of those moderators in Model 1.

The R^2 of Model 2 was .59, which indicates that about 59% of total between-study variance was explained by Model 2. Although this model had a limited number of studies ($n = 35$) compared to the previous model (Model 1), this model explained an additional 31% of between-study variance.

Table 28

Meta-Regression Results for Model 2 (All Moderators, n = 35)

Covariate	Coefficient	SE	95% C.I. [low, high]	*Z*	*p*	*p* as a set
Intercept	1.987	.695	[.624, 3.349]	2.86	.004	-
Type of publication (Journal vs. Dissertation)	-.684	.139	[-.956, -.412]	-4.93	.000	.000
Type of publication (Journal vs. Tech report)	-.085	.051	[-.184, -.014]	-1.69	.092	
Grade level difference (Primary vs. Intermediate)	-.071	.035	[-.139, -.003]	-2.03	.042	.000
Grade level difference (Primary vs. secondary)	-.218	.045	[-.306, -.13]	-4.86	.000	
Female %	-2.753	1.217	[-5.139, -.368]	-2.26	.024	-
ELL %	.029	.333	[-.623, .681]	.09	.930	-
SpEd %	2.159	.759	[.669, 3.647]	2.84	.005	-
FRL %	-.568	.241	[-1.04, -.095]	-2.35	.019	-
White %	-.144	.161	[-.459, .171]	-.09	.370	-
Type of CBM (oral reading vs. maze)	.080	.067	[-.05, .211]	1.2	.229	-
Type of CBM passage (commercial vs. researcher-developed)	.653	.162	[.335, .971]	4.02	.000	-
Number of CBM passage (one vs. two)	.009	.097	[-.182, .199]	.09	.929	.000
Number of CBM passage (one vs. three)	.235	.047	[.143, .328]	4.99	.000	
Type of CBM administrator (Researcher vs. school personnel)	-.139	.09	[-.316, .038]	-1.54	.123	.003
Type of CBM administrator (Researcher vs. N/A)	.177	.095	[-.009, .363]	1.86	.063	
Type of state test (commercial vs. state-developed)	.181	.078	[.028, .333]	2.33	.02	-
Response format of state test (Multiple choice vs. mixed format)	.207	.071	[.068, .347]	2.91	.004	-
Time interval between CBM and state test	.028	.014	[.001, .055]	2.01	.045	-

Model Statistics					
Test of the model			Q	df	*p*-value
			190.55	18	.000
Goodness of fit	T^2	I^2(%)	Q	df	*p*-value
	.002	97.84	740.86	16	.000
Total between-study variance	T^2	I^2(%)	Q	df	*p*-value
	.0049	99.33	5077.45	34	.000
Proportion of total between-study variance explained by model (R^2)				.59	

Note. N/A for type of CBM administer indicates correlations with no information.

A comparison of moderator effects between Model 1 (all moderators without student demographics) and Model 2 (all moderators including student demographics) is summarized in Table 29.

Table 29

Comparison of Moderator Effects (p-value) between Model 1 and Mode 2

Moderator	Model 1: Moderators except demographics (*n*=129)	Model 2: All moderators (*n*=35)
Type of publication (Journal/Dissertation/Tech report)	.025	.000
Grade level difference (Primary/Intermediate/Secondary)	.008	.000
Female %	Not included	.024
ELL %	Not included	.930
SpEd %	Not included	.005
FRL %	Not included	.019
White %	Not included	.370
Type of CBM (Oral reading/Maze)	.029	.229
Type of CBM passage (Commercial/Researcher-developed)	.657	.000
Number of CBM passage (1/2/3)	.366	.000
Type of CBM administrator (Researcher/School personnel)	.167	.003
Type of state test (Commercial/State-developed)	.001	.02
Response format of state test (Multiple choice/Mixed format)	.729	.004
Time interval between CBM and state tests	.000	.045

Publication Bias

To analyze possible publication bias, which indicates a tendency for studies to be more likely to be published when they have larger effects, two funnel plots were examined. The first funnel plot in Figure 3 depicts correlations against precision. Large studies were located around the top of the plot, and smaller studies dispersed across a range of values at the bottom of the plot due to more random variation in the small studies (Borenstein et al., 2009). As shown in the plot, most small studies were distributed quite symmetrically about the estimated mean correlation, which indicates no evidence of publication bias. If there was publication bias, the bottom side of the plot would have shown a higher concentration of studies on one side about the mean (the middle line) than the other. Figure 4 shows correlations plotted against standard error, the inverse of precision, and also showed no tendency of publication bias.

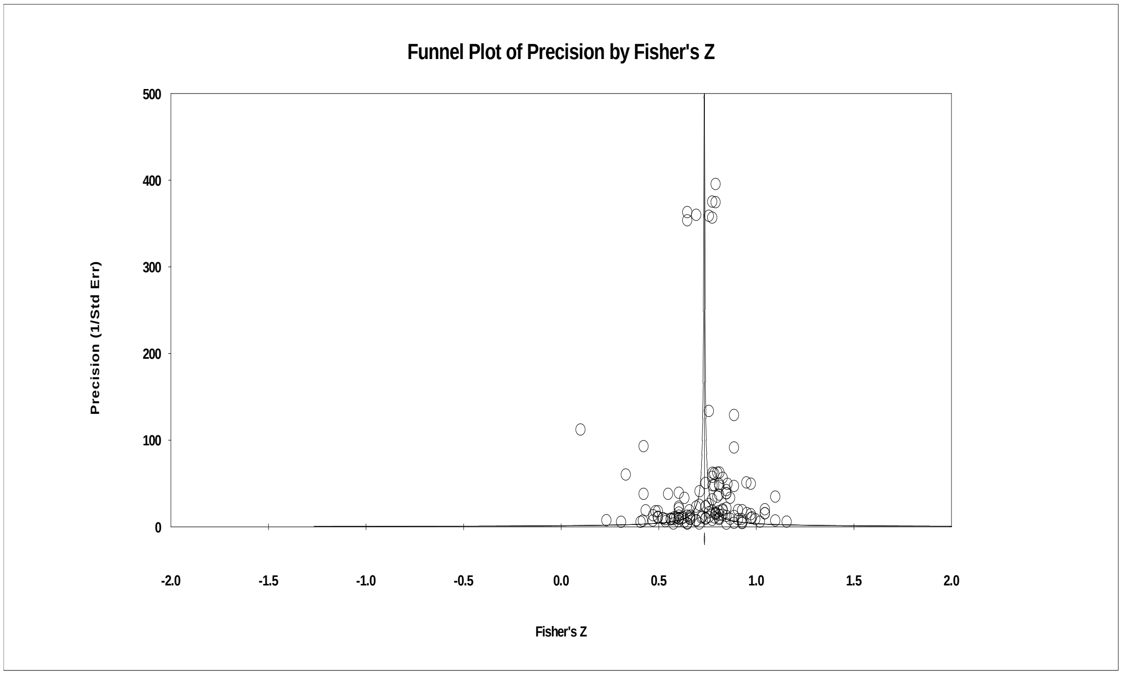


Figure 3. Plot of Correlations against Precision

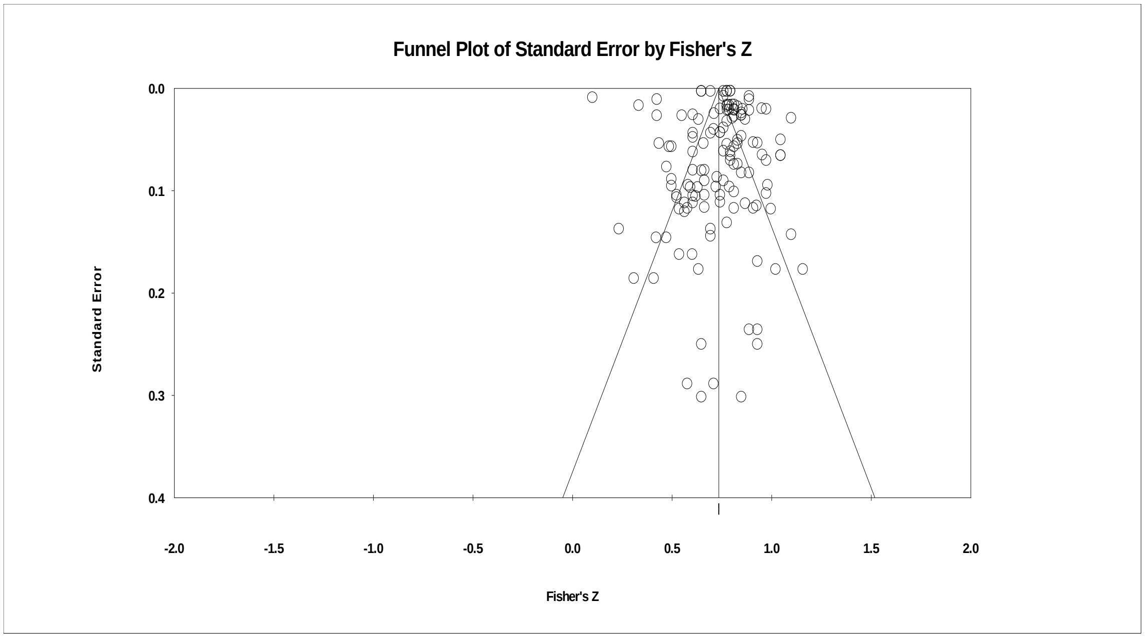


Figure 4. Plot of Correlations against Standard Errors

CHAPTER V

DISCUSSION

The purpose of this study was to examine the extent to which CBM oral reading and maze predict performance on reading comprehension measured by state achievement tests across grades. Given that relatively less is known about the validity of oral reading and maze in relation to reading comprehension than to overall reading proficiency, the present study sought to quantitatively synthesize findings on the validity of the two CBMs in relation to reading comprehension specifically measured by state achievement tests using a meta-analysis. To address this purpose, predictive validity of oral reading and maze was investigated across primary, intermediate, and secondary grades, and the effects of potential moderators were investigated. Sixty-one studies were identified and a total of 132 correlations were used for the meta-analysis.

In this section, I discuss (a) what the current study results indicate in terms of validity of oral reading and maze in relation to reading comprehension measured by state tests and factors (moderators) that influenced the relations, especially in light of previous theoretical and empirical literature, (b) implications of the results for research and practice, and (c) limitations and directions for future research.

RQ1: What are the estimated average correlations between CBM tasks (oral reading and maze) and reading comprehension on state achievement tests, and do they vary by (a) type of task (oral reading vs. maze) and (b) grade range (primary, intermediate, secondary)?

The first main research question was to estimate the average correlation derived from the literature and to see whether the strength of the relations differ by oral reading or maze and by grade level. The first hypothesis was that, for primary grades (Grades 1-3), the estimated average correlation between oral reading and reading comprehension would be larger than that between maze and reading comprehension. The second hypothesis was that, for intermediate and secondary grades (after Grade 4), estimated average correlation between maze and reading comprehension would be larger than that between oral reading and reading comprehension. The first hypothesis was confirmed; however, the second hypothesis was not supported: there was a significant main effect of type of CBM, favoring oral reading across grade levels.

Specifically, the overall strength of the relations between CBM (oral reading and maze) and reading comprehension measured by state achievement tests was *r* = .63, which was interpreted as large correlation based on Cohen's (1992) criteria– small (*r* = .1), medium (*r* = .3), and large (*r* = .5). This finding corroborates other researchers' conclusions that CBM reading (oral reading and maze task) function as valid indicators of reading comprehension on state tests across grades (e.g., Graney et al., 2010; Wayman et al., 2007). The mean correlation (*r* = .63) for the present meta-analysis was slightly lower than the reported correlation in a recent meta-analysis (*r* = .69 for Yeo, 2010). While this difference does not seem to be large from a practical perspective, one possible reason for this slight difference may be that Yeo's (2010) study estimated the correlation between CBM and overall reading achievement on state tests whereas the present study estimated the correlation between CBM and reading comprehension, specifically. It might be possible that CBM oral reading and maze are slightly less predictive of specific reading comprehension than overall reading achievement. Another reason may be that Yeo's (2010) study included only three studies on maze, while the present meta-analysis included 14 studies, which might provide a better estimate of the correlation.

According to the type of CBM, the mean correlation between oral reading and state tests of reading comprehension across grades significantly differed from that of maze (p = .029), when other covariates were held constant. Specifically, the mean correlation for oral reading (r = .63) was larger than that for maze (r = .60). This result indicates that oral reading appears to be slightly more predictive of reading comprehension performance on state tests than maze does when other covariates held constant. This result, however, provides only overall mean correlations of both CBM tasks when all grades are taken together; thus, it needs to be further examined to see how both CBMs function for each grade range (primary, intermediate, and secondary).

Regarding the effect of grade range, there was a significant grade range effect (p = .008) on the strength of the relations between CBM and state tests of reading comprehension when other covariates were held constant. Specifically, the mean correlation for primary grades (1-3) was r = .63, which was slightly lower than the mean correlation for intermediate grades (r = .65). The lowest correlation was found for secondary grades (r = .60). The mean correlation did not significantly differ between primary and intermediate grades (p = .954), but there was a significant difference between primary and secondary grades (p = .007). These results indicate that CBM reading tasks (oral reading and maze) have evidence of similar validity across elementary grades (i.e., grades 1-6); however, the strength of these relations decreased at the secondary (middle and high school) level.

Next, to further investigate the grade range difference for each CBM task, a series of moderator analyses were conducted for oral reading and maze separately. For oral reading, the mean correlation was highest for the intermediate grades and lowest for secondary grades. This difference, however, was not significant (Q = .513, *df* = 2, *p* = .773), suggesting that the mean correlation did not significantly differ by grade range. A similar correlation pattern was observed for maze, with the highest correlation for intermediate and the lowest correlation for secondary grades. However, unlike oral reading, this difference was significant (Q = 11.34, *df* = 2, *p* = .003), which indicates that the mean correlation for maze differed by grade range. Moreover, when the analyses were conducted for primary, intermediate, and secondary grades separately, the average correlations for oral reading were larger than those for maze for all grade ranges, although there were not significant differences between oral reading and maze.

To summarize, the meta-regression analyses revealed that the average correlation for CBM oral reading was significantly larger than that for maze when all grade ranges were taken together, and the average correlation for oral reading was larger than that for maze at each grade range (primary, intermediate, secondary), although the differences were not significant. Given that the analyses for the effect of type of CBM in each grade range were conducted separately (thus using lower sample sizes), non-significant differences between oral reading and maze might be due to lower power. However, when combining intermediate and secondary levels together (to align with hypothesis 2), a significant difference between the average correlations (*r* = .64 for oral reading and .59 for maze) was detected (*p* < .01), which indicates that oral reading may be more predictive of reading comprehension than maze for students after Grade 4. In addition, the average correlation for maze decreased considerably (*r* = .64 to .57) between intermediate and secondary grades, whereas the average correlations for oral reading were relatively stable across grade range without significant difference by grade range. These stable relations between oral reading and states tests of reading comprehension corroborate previous meta-analytic findings that demonstrated consistent relations across elementary grades (Reschly et al., 2009).

Results from the current meta-analysis, however, contradict the second hypothesis that maze would be more predictive of reading comprehension than oral reading for intermediate and secondary grades. Further, findings are somewhat surprising in that previous researchers have noted that correlations between oral reading and general reading proficiency on criterion measures tend to be stronger at the primary grades and decrease at the intermediate grades, whereas correlations for maze remain fairly stable across the grades (e.g., Graney et al., 2010; Jenkins & Jewell, 1993; Wayman et al., 2007). This expected pattern can be explained in part by developmental reading theory; once decoding is mastered, oral reading fluency no longer accounts for significant variability in reading proficiency (Marcotte & Hintze, 2009). Rather, as the primary emphasis of reading instruction switches from decoding and fluency to comprehension after Grade 3, other factors (e.g., vocabulary, inference skills) become increasingly important to the relation with reading proficiency (Cain & Oakhill, 1999; Graney et al., 2010).

Although previous literature did not provide strong empirical evidence that maze is superior to oral reading for assessing reading comprehension (Graney et al., 2010), some previous researchers suggested that maze may measure reading comprehension more directly than does oral reading (Tolar, Barth, Francis, Fletcher, Stuebing, & Vaughn, 2011). That is, researchers have argued that performance on maze is related to vocabulary, background knowledge, syntactic skills, or inference-making competency, which explain unique variance of reading comprehension, whereas oral reading cannot (e.g., Catts et al., 1999; Tolar et al., 2011). Further, maze has been known to have greater face validity as a comprehension measure than oral reading among teachers (Fuchs, Fuchs, & Maxwell, 1988; Wayman et al., 2007).

In light of the above arguments, although average correlations for both oral reading and maze were “large” based on Cohen’s standards, the larger correlation for oral reading compared to maze across grade levels in this meta-analysis is somewhat unexpected. Although oral reading has been criticized for emphasizing one aspect of the simple view of reading, that is, decoding (Graney et al., 2010; Munger & Blachman, 2013; Wayman et al., 2007; Yeo, 2010), results of the present study strengthen the empirical evidence that oral reading fluency is a strong indicator of reading comprehension across grades. Furthermore, findings of this meta-analysis suggest that oral reading might be a better indicator of reading comprehension than maze even for intermediate or secondary grade levels, which contradicts previous findings (e.g., Wayman et al., 2007).

Given that the role of decoding skills decreases as students' grade increases (Catts et al., 1999; Catts et al., 2003; Graney et al., 2010), more theoretical as well as empirical investigation is warranted to clarify the mechanism of how students' oral reading proficiency more closely relates to reading comprehension than maze does across grades. Perhaps, the automatic information processing theory suggested by LaBerge and Samuels (1974) would be most suitable to explain the close association between oral reading fluency and reading comprehension. That is, once automatic decoding is secured, more cognitive resources can be used to the processing of meaning of the text (Cutting & Scarborough, 2006; LaBerge & Samuels, 1974). This way, oral reading fluency can function as an indicator not only of decoding skill but also of comprehension of the text (Fuchs, Fuchs, Hosp, & Jenkins, 2001).

Another possible explanation would be that oral reading and maze may differ regarding the sensitivity of each measure to capture small differences between students. That is, a student's performance on the oral reading task is measured by counting every word the student read, whereas for maze, relatively fewer number of words actually count toward the score, and thus may not be as sensitive to slight differences among readers. In other words, there might be a plateau for maze, which might restrict opportunity for a range in responses at the higher end. In addition, for maze, researchers have pointed out that maze may only measure sentence level comprehension, rather than paragraph or passage (discourse) level comprehension (Carlson, Seipel, & McMaster, 2014; January & Ardoin, 2012; Parker, Hasbrouck, & Tindal, 1992). For instance, January and Ardoin (2012) found that, although context may help students completing the maze, context beyond the sentence level was not needed for selecting 90% of the target words accurately. In this respect, it might be possible that maze does not reliably estimate reading comprehension as previous research suggested (Parker et al., 1992).

Thus far, I have discussed the effects of type of CBM and grade level difference on the strength of the relations between CBM and reading comprehension on state tests. Given that the moderator analyses indicated that significant variance remained after either type of CBM or grade level difference was used as a covariate, more explanation about the role of other moderating variables is required.

RQ2: To what extent are the relations between CBM tasks (oral reading and maze) and reading comprehension on state achievement tests influenced by potential moderating factors?

In addition to the two significant moderators discussed above (effect of type of CBM and grade level difference), the following three moderators were significant sources of variance in correlations: type of publication (journal articles, dissertations, technical reports), development type of state test (commercial or state-developed), and time interval between CBM and state test administration.

Specifically, regarding type of publication, correlations from journal articles were significantly higher than those from dissertations, whereas there was not a significant difference in correlations between journal articles and technical reports. Given that studies with relatively small effects (i.e., lower correlations) would be less likely to be published, it makes sense that correlations from published journal articles were larger than those from dissertations. It was interesting, however, that correlations from technical reports reported higher correlations between CBM and reading comprehension on state tests than journal articles did. One possible explanation is that the majority of technical reports included in the present meta-analysis were conducted for oral reading which generally yielded higher correlations than maze.

In terms of the effect of development type of state test, correlations from state-developed tests were significantly higher than those from commercially prepared tests of reading comprehension. This result is unexpected given that commercial criterion measures are often assumed to be more technically sound (Reschly et al., 2009). In addition, state-developed tests tend to be heterogeneous in terms of content, format, and difficulty level due to different curricula and standards (Peterson & Hess, 2005; Reschly et al., 2009). One possible explanation would be that only 22 correlations from 10 studies using commercially prepared tests (e.g., ITBS, SAT-10, TerraNova) were included in the meta-analysis, whereas the majority of studies reported correlations with state-developed tests. This result demonstrates that CBM reading tasks can be used as valid predictors of performance on state-developed tests of reading comprehension.

Regarding the effect of time interval, correlations were higher when the time interval between CBM and state test administration was shorter. This result corresponds to previous studies (e.g., Reschly et al., 2009; Roehrig et al., 2008; Yeo, 2010) that showed that the strength of correlations tended to decrease when the time interval between the measurement occasions increased. It makes sense that correlations would be stronger with shorter time intervals in that there is less time for instruction and other intervening factors to affect student performance. Although some studies demonstrated that the time interval between CBM and state tests (CBM administered in fall, winter, and spring) did not influence the correlations for Grade 4 students (Shapiro et al., 2006), the present meta-analysis along with previous meta-analyses strengthen the notion that the shorter time interval between CBM and state tests is likely to produce the higher correlations. Of course, it should be noted that many other sources of variations across studies could influence these relations.

Other moderators, however, did not have significant effects on the relations between CBM and reading comprehension on state tests. Non-significant moderators included development type of CBM passage, number of CBM passages used, type of CBM administrator, and response format of state tests.

Specifically, correlations did not significantly differ according to the development type of CBM passage. That is, commercially prepared CBM or researcher-developed CBM passages produced statistically non-significant average correlations with reading comprehension on state tests. Although the average correlation for researcher-developed CBM was slightly higher than that for commercial CBM passages, this result indicates that the two types of CBM passages are likely to produce similar predictive validity coefficients. In addition, this result is consistent with results of previous studies (Fuchs & Deno, 1992; Yeo, 2008; 2010). However, given that the majority of studies (about 84%) in this meta-analysis used commercially prepared standardized passages, further research is warranted to examine the effect of the nature of CBM passages.

Next, correlations did not significantly differ by the number of CBM passages used. This result suggests that the data used to comprise the final score (i.e., a single score from one passage, a mean score from two, or a mean or median score from three passages) did not influence the relations between CBM and reading comprehension on state tests. This result is consistent with previous meta-analysis (Yeo, 2008), and it also supports Ardoin et al. (2004) study that demonstrated using a single passage might be sufficient for screening purposes. Given that more than half of the studies used three CBM passages to yield a single score, this result suggests a more efficient way of using CBM as screening measures in research and practice.

Similarly, type of administrator of CBM (researchers or school personnel) did not influence the relations between CBM and reading comprehension on state tests. Correlations for both type of administrators were very similar, meaning that the validity of CBM (oral reading and maze) as an indicator of reading comprehension on state tests would be the same regardless of the type of administrator. These results support the argument that CBM reading tasks can be used flexibly across different conditions, such as materials and administrators (Wayman et al., 2007; Yeo, 2008).

Further, correlations did not significantly differ by response format. Although the average correlation for mixed (i.e., multiple choice plus open ended questions) format was slightly higher than that for multiple choice format of state tests, there was statistically significant difference between them. Given that the nature of criterion measures has been considered potential moderators influencing the relations between CBM and criterion measures (e.g., Reschly et al., 2009; Wayman et al., 2007), this result contradicts previous findings. One possible explanation would be that state tests with mixed formats included in this study were not very different from other state tests with multiple choice format in terms of difficulty or processing demands. However, because the present meta-analysis did not provide a detailed examination of response format, further research is needed on the effect of response format of criterion measures.

In terms of the effects of student characteristics (based on Model 2, *n* = 35 correlations), proportion of females, special education, and FRL were significant moderating variables. For the effect of female proportion, as one unit of female proportion increased in the sample, the correlation between CBM and reading comprehension on state tests decreased by -2.75. This result contradicts previous findings, given that Yeo (2010) reported no moderating effect of gender. Therefore, more research is needed to further investigate whether CBM reading tasks are less predictive of reading comprehension on state tests for female students.

In contrast, the proportion of students receiving special education in the sample had a positive effect on the correlations, indicating that studies with a higher proportion of students in special education produced higher correlations between CBM and state tests of reading comprehension. The positive effect of special education proportion can be explained in relation to the sample size of studies. That is, studies with a large sample size generally had high correlations, and those studies with a large sample size tended to have higher proportions of students in special education. However, given that the previous study (e.g., Yeo, 2010) found that studies with high proportions of students in special education yielded lower correlations than did studies with lower proportion of students with disabilities, more investigation is warranted.

Regarding FRL, studies that included a higher proportion of students on FRL produced lower correlations between CBM and reading comprehension on state tests. This result is consistent with Hixson and McGlinchey's (2004) study that demonstrated FRL status could be a source of significant contribution to predicting performance on criterion tests in reading. In addition, this result suggests that the two CBM reading tasks may be less predictive of reading comprehension on state tests as study samples include more students who receive FRL. One possible reason would be that higher proportions of students receiving FRL might lead to less variability in the sample because students receiving FRL are more likely to have lower reading performance (e.g., Hixon & McGlinchey, 2004) which may yield a restricted range of scores, and that less variability in the sample would lead to lower correlations (Goodwin & Reach, 2006). However, more research is needed before one can be sure whether these student demographics (female, special education, FRL) are important factors as practitioners use oral reading and maze as indicators of reading comprehension, given these results were based on very limited sample size compared to the whole sample.

Other demographic variables (proportion of ELL and White students, which served as a proxy for diversity in the sample) were not significant moderators, meaning that that proportion of ELL and White students in the sample did not influence the correlations between CBM and reading comprehension measured by state tests. This result suggests that students' language and ethnicity backgrounds may not influence the validity of CBM reading tasks; however, more research is warranted to better understand the role of language and ethnicity backgrounds.

Regarding the effects of student demographics discussed thus far, it should be noted that, compared to Model 1 which did not include student demographic information, Model 2 involved only 35 correlations because there were many missing values for demographic information. Thus, results regarding effects of student demographics derived from the Model 2 should be interpreted very cautiously. Given that this issue has been a problem in previous meta-analyses as well (e.g., Reschly et al., 2009), where there were not sufficient data to even include demographic information in the analyses, further research is needed for a more thorough examination of the effects of student demographics on the relations between CBM and reading comprehension on state tests.

Limitations and Directions for Future Research

Although the present study conducted a thorough quantitative review by a meta-analysis, several limitations clearly exist.

First, in this meta-analysis, only one correlation for each CBM task was selected per study to avoid the problem of dependency, which inflates the available information and overestimates confidence in the results (Van den Noortgate et al., 2013). To address the dependency issue, researchers can either average multiple correlations or choose one correlation that is most of interest. However, simply averaging correlations within studies could lead to the artificially reduced variance between correlations, and therefore, informative differences between correlations could be lost (Cheung & Chan, 2008). The present meta-analysis adopted the second option (i.e., choosing one correlation that is most of interest) by picking one correlation representing the most concurrent relation between CBM and state tests. Nevertheless, it should be noted that picking one correlation per study might not perfectly reflect the true relations.

Second, the grade-level moderator had to be categorized according to grade *range* (primary, intermediate, and secondary) due to very small sample sizes for some grades (e.g., grades 9-10). As mentioned earlier, the small number of correlations per category could lead to biased estimates by overestimating within-group variance (Borenstein et al., 2009). Although the "grade range" categorization has a practical implication in terms of the use of CBM reading tasks as a valid indicator of reading comprehension for primary, intermediate, and secondary grade levels, future research is warranted for further investigation on the validity of CBM reading tasks for each grade level.

Third, regarding the effects of student demographics, a limited number of study correlations (only 35 out of 132 correlations) were included for Model 2, which involved all moderators including student demographics) due to missing values. For this reason, Model 2 should be interpreted with caution, although Model 2 sought to reflect all moderating variables' effects at the same time. Given that previous studies suggest that student demographics might influence the relations between CBM and criterion measures of reading (e.g., Wayman et al., 2007; Yeo, 2008; 2010), more investigation will be needed for analyzing the effects of diverse student demographic information. To avoid considerable number of missing values, researchers should report student demographic information in their studies. Another limitation regarding the student demographics is that the proportion of White students was used as a proxy for diversity in the sample. Although the proportion of White students was needed for the purpose of the meta-analysis, it should be noted that this variable might be problematic in that students of other races are lumped into one category.

Fourth, in this meta-analysis, some moderators (e.g., type of CBM scores [static or slope score], norm- versus criterion-referenced state tests) that have been investigated in previous studies were not included. For the type of CBM scores, given that only a few studies examined the relations between CBM slopes and reading comprehension, it was not possible to investigate the effect of the type of CBM scores. As Fuchs (2004) demonstrated, however, slope on CBM tasks is a crucial element in examining a student's responsiveness to instruction; therefore, more research is needed to investigate whether slope is a valid indicator of reading comprehension across grades. In addition, regarding the difference between norm- or criterion-referenced state tests, the moderator was nearly identical to the developmental type of state tests (commercial or state-developed) because most state-developed tests were criterion-referenced. Given that the meta-regression model (Model 1) explained only 28% of the variance, more investigation is warranted on other potential moderating factors.

In addition to the limitations discussed so far, a few future directions emerge from this meta-analysis. First, the overwhelming majority of research in CBM as an indicator of reading comprehension has been conducted with oral reading; a limited number of studies (29 correlations out of 123 correlations; 14 studies out of 61 studies) focused on maze in this meta-analysis. Thus, further research is warranted to provide more empirical support and guidance in using maze as a valid indicator of reading comprehension across primary to secondary grades. Specifically, more empirical evidence is needed to support the notion that maze might be a better indicator of reading comprehension for intermediate and secondary students than is oral reading (Jenkins & Jewell, 1993; Wayman et al., 2007). In terms of theory, more research is required to determine whether maze assesses reading comprehension abilities depending on the context, beyond merely sentence meaning and structural repetition (Carlson et al., 2014; Parker et al, 1992).

Second, a closer examination should be given to oral reading as a screening tool for detecting reading comprehension difficulties, particularly for the group of students often called “word callers.” These students are known as those who appear to read fluently (i.e., they have intact phonological or decoding skills) but do not comprehend well (Hamilton & Shinn, 2003). It is estimated that about 10-25% of poor readers have this profile, especially in the intermediate grades or older students (Catts et al., 2004; Cutting & Scarborough, 2006). Although the present meta-analysis revealed that oral reading is generally a better indicator of reading comprehension than maze across grades, more research is needed to examine whether those word-calling students can be accurately identified through oral reading task and whether a combination of oral reading and other indicators of reading comprehension may improve the identification accuracy.

Third, more research is needed on the validity of oral reading and maze as a technically valid indicator of reading comprehension for students at the secondary level. Many studies (92 correlations out of 132 correlations) have focused on primary and intermediate grades, whereas relatively few studies (30 correlations out of 132 correlations) have examined the validity of oral reading and maze for secondary students.

For establishing a seamless and flexible system of CBM in reading (Wayman et al., 2007), more empirical evidence on secondary levels should be further accumulated.

Fourth, more research is needed on the validity of oral reading and maze for various populations with diverse backgrounds (e.g., language, ethnicity, SES, gender, disabilities). Examining the extent to which scores of CBM function similarly for students with various backgrounds is paramount to establishing the pertinence of inferences drawn from CBM tasks (Reschly et al., 2009). Given that a limited number of studies in this meta-analysis provided students' demographic information, further efforts are required to see whether there are any potential effects of student characteristics on the relations between CBM reading tasks and reading comprehension on state tests. For example, if the oral reading task appears to overestimate the performance of a certain population, such as African American students (Kranzler et al., 1999), it could result in under-identification for services. In practice, teachers may need substantiated evidence that indicates CBM as useful and unbiased measures for an increasingly diverse body of students who are at risk for reading comprehension (McMaster et al., 2006).

Fifth, it should be noted that, for oral reading and maze to be used as a screening tool to specifically identify students with reading comprehension difficulties, more evidence is needed about classification accuracy (Hintze & Silberglitt, 2005; Jenkins, Hudson, & Johnson, 2007), given that the validity evidence is necessary, but not sufficient to validate CBM for screening. Although the average correlation in this meta-analysis indicates relatively strong criterion-related validity, there might be a potential of misclassification rates in high-stakes decision. Thus, more research is needed for improving classification accuracy (e.g., sensitivity and specificity) of CBM. For example, further evidence of whether CBM oral reading and maze can predict performance on state tests accurately (i.e., pass or fail) will support using CBM as a screening tool for reading comprehension difficulties (Graney et al., 2010). Given that the state tests might vary and not be comparable in terms of the content, performance standard, difficulty, more research is needed to examine whether CBM oral reading and maze yield good sensitivity and specificity for different state tests (Jenkins et al., 2007).

Related to the issue of CBM as an effective screening tool, multiple-gating approaches in screening procedures should be considered and further investigated. That is, more research is needed on the value of combining multiple CBM measures for improving predictive validity and classification accuracy for at-risk students (Jenkins et al., 2007). There is some evidence that more than one screening measure yielded improved classification accuracy (O'Connor & Jenkins, 1999) or combining static assessment (a one-time screening) and subsequent progress monitoring in a short period (e.g., 5 weeks) increased sensitivity and specificity (Fuchs et al., 2004). However, cost-benefit trade-offs should be considered when employing multiple-gating approach in practice because more cost and time for additional administration would be needed.

Implications for Practice

Despite the limitations of the present meta-analysis and need for further research, findings of this meta-analysis provide several implications for current practice. First, given that the present study quantitatively synthesized the existing research base rather than a single study result, this study provides practitioners with converging evidence about the concurrent and predictive validity of CBM oral reading and maze (Rosenthal et al., 2006). Thus, integrative conclusions drawn from this meta-analysis suggest that CBM oral reading and maze can be used as a valid indicator to assess students' reading comprehension proficiency across grades.

However, the validity evidence differed by the type of CBM reading task and grade levels. That is, the strength of the relations between CBM reading tasks and reading comprehension for primary grades did not significantly differ from that for intermediate grades; however, the strength of the relations for secondary grades was significantly weaker than that for primary or intermediate grades. Therefore, it can be said that CBM oral reading and maze might be used as better indicators to assess reading comprehension for elementary school students than for middle or high school students. Further, the validity of oral reading and maze for primary grades is expected to be similar to that for intermediate graders; therefore, teachers can use both oral reading and maze task across elementary grades, but should keep in mind the decreased validity for secondary students.

In terms of the effect of type of CBM task, the present meta-analysis suggests that oral reading task is a slightly better indicator to assess students' reading comprehension across grades. Specifically, although the validity of oral reading and maze did not significantly differ for each grade range (primary, intermediate, and secondary), oral reading showed larger estimated mean correlations with reading comprehension on state tests than did maze across grades. However, given that the average correlations for oral reading and maze were both "large" by Cohen's (1992) standard and that maze is more time-efficient by group-administering, results suggest that practitioners might use oral reading and maze interchangeably for students from elementary to secondary grades, whereas more evidence is needed for secondary level students.

In addition, the current meta-analysis' results for oral reading corroborate the evidence that oral reading can be used as a valid indicator or predictor of overall reading competence, including reading comprehension (Baker et al., 2009; Fuchs et al., 2001; Thurlow & van den Broek, 1997). Further, although it has been reported that teachers tend to be reluctant to use CBM oral reading as a reading comprehension measure because, in part, oral reading does not seem to require students to understand the text (Fuchs et al., 2001; Jenkins & Jewell, 1993; Wayman et al., 2007), the present meta-analysis demonstrates that teachers may have confidence when they use oral reading as an indicator of reading comprehension across grades. Nevertheless, more evidence is needed at the individual level, such as the validity of oral reading for word-calling students (Wayman et al., 2007).

Regarding the use of CBM reading tasks across various settings (e.g., different administrators, materials, and so on), results of the present meta-analysis support the view that CBM oral reading and maze involves flexibility and consistency across different settings (Wayman et al., 2007). Specifically, the validity of CBM reading tasks did not differ by development type of CBM passage, number of CBM passages used, type of CBM administrator. These results suggest that, in practice, teachers can administer CBM reading tasks in a reliable manner, and both commercially- and researcher-developed CBM passages can be used to predict state tests of reading comprehension. In addition, this study indicates that teachers may administer one or two passages instead of three for efficiency when assessing students' reading comprehension, given that the validity of CBM reading tasks was constant regardless of the number of CBM passages used. It should be noted, however, that this result may not hold true when using CBM reading tasks for purposes of progress monitoring.

Finally, in terms of using CBM oral reading and maze for students with various characteristics, there were significant student characteristic effects (proportion of female, special education, and FRL) on the relation between CBM and reading comprehension on state tests. It means that those student characteristics might lead to a significant contribution to predicting reading comprehension on state tests, and therefore, practitioners might need to use caution when using CBM oral reading and maze to predict reading comprehension on state tests. However, given that this meta-analysis was not able to conduct a more thorough examination of the moderating effects of student characteristics due to too many missing data, future research is needed to provide more clear evidence.

Conclusion

According to Messick (1989b), the validity of a measure can be established by an ongoing and recursive process; therefore, it is not determined by one study but by the body of evidence on the validity accumulated over time. In this respect, the present meta-analysis sought to provide a more comprehensive examination to support the validity of two widely used CBM reading tasks in relation to reading comprehension measured by state achievement tests, with a focus on differential validity between oral reading and maze for different grade levels of students. Additionally, potential moderating factors (characteristics of students, CBM, and state tests) were analyzed to investigate impacts of different educational factors.

The validity of the two CBM reading tasks as screening tools for identifying students with reading comprehension difficulties seems to be promising across grades, with oral reading showing slightly stronger relations to reading comprehension across grade levels. In addition, when considering the fact that each state test was built to measure reading competency based on their own academic standards, CBM oral reading and maze appears to be a valid predictor of reading comprehension (Shapiro et al., 2006; Wayman et al., 2007). Further in-depth research is warranted for investigating the validity of CBM reading tasks for secondary students and students with diverse backgrounds (e.g., special education, FRL status). As CBM tasks are increasingly used to predict performance on state achievement tests (Denton et al., 2011; Graney et al., 2010), those ongoing efforts will shed light on establishing a seamless and flexible system of CBM for identifying students with reading comprehension difficulties.

References

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

Acquavita, T. L. (2012). *A longitudinal exploration of the relationship between oral reading fluency and reading comprehension achievement among a sample of diverse young learners* (Unpublished doctoral dissertation). Florida International University, Miami: FL.

*Alonzo, J., & Tindal, G. (2004). *Analysis of Reading Fluency and Comprehension Measures for Sixth Grade Students* (Technical Report #24). Eugene, OR: Behavioral Research and Teaching.

*Anderson, D., Alonzo, J., & Tindal, G. (2011). *easyCBM Reading Criterion Related Validity Evidence: Oregon State Test 2009-2010* (Technical Report #1103). Eugene, OR: Behavioral Research and Teaching.

*Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., & Williams, K. L. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33*(2), 218-233.

*Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., & Beck, C. T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review, 37*(1), 18.

*Baker, D. L., Biancarosa, G., Park, B. J., Bousselot, T., Smith, J., Baker, S., Kame'enui, E. J., Alonzo, J., & Tindal, G. (2015). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in Grades 7 and 8. *Reading and Writing, 28*(1), 57-104.

Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end of grade reading assessment* (Technical Report). Asheville, NC: North Carolina Teacher Academy.

Bishop, D. V., & Snowling, M. J. (2004). Developmental dyslexia and specific language impairment: Same or different? *Psychological Bulletin, 130*, 858-886.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Method, 1,* 97-111.

*Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report #1). Tallahassee, FL: Florida Center for Reading Research.

Cain, K., & Oakhill, J. (2007). Reading comprehension difficulties: Correlates, causes, and consequences. In K. Cain & J. Oakhill (Eds.), *Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 41-76). New York, NY: Guilford.

Cain, K., Oakhill, J., & Bryant, P. (2000). Phonological skills and comprehension failure: A test of the phonological processing deficit hypothesis. *Reading and Writing: An Interdisciplinary Journal, 13,* 31-56.

*Canto, A. I. (2006). *Predicting third grade students' FCAT reading achievement and oral reading fluency using student demographic, academic history, and performance indicators* (unpublished doctoral dissertation). Florida State University, Tallahassee: FL.

Catts, H. W., Adlof, S. M., & Weismer, S. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49*, 278-293.

Catts, H. W., Herrera, S., Nielsen, D. C., & Bridges, M. S. (2015). Early prediction of reading comprehension within the simple view framework. *Reading and Writing*, *28*, 1407-1425.

Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences, 18,* 329-337.

*Conway Sledge-Murphy, F. (2011). *The Relationship between Oral Reading Fluency and Reading Comprehension for Third Grade Students in a Rural Louisiana School District* (Unpublished doctoral dissertation). University of Louisiana at Monroe, Monroe: LA.

*Cook, R. G. (2003). *The utility of DIBELS as a curriculum based measurement in relation to reading proficiency on high stakes tests* (Unpublished master's thesis). Marshall University, Huntington: WV.

*Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide assessment tests. *Educational Assessment, 7*, 303-323.

*Decker, D. M., Hixson, M. D., Shaw, A., & Johnson, G. (2014). Classification accuracy of oral reading fluency and maze in predicting performance on large-scale reading assessments. *Psychology in the Schools, 51*, 625-635.

DeCoster, J. (2005). Meta-analysis. In K. Kempf-Leonard (Ed.), *The Encyclopedia of Social Measurement*, 683-688. San Diego, CA: Academic Press.

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184-192.

*Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., Romain, M., & Francis, D. J. (2011). The relations among oral and silent reading fluency and comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading, 15*, 109-135.

*Devena, S. (2013). *Relationship of oral reading fluency probes on students' reading achievement test scores* (Unpublished doctoral dissertation). Arizona State University, Tempe: AZ.

*Echols, J. M. Y. (2010). *The utility of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in predicting reading achievement* (Unpublished doctoral dissertation). Seattle Pacific University. Seattle: WA.

Elleman, A. M., Compton, D. L., Fuchs, D., Fuchs, L. S., & Bouton, B. (2011). Exploring dynamic assessment as a means of identifying children at risk of developing comprehension difficulties. *Journal of Learning Disabilities, 44*, 348-357.

Espin, C. A., Deno, S. L., Maruyama, G., & Cohen, C. (1989, March). *The Basic Academic Skills Samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms*. Paper presented at the National Convention of the American Educational Research Association, San Francisco, CA.

*Espin, C., Deno, S., McMaster, L., Wayman, M., Yeo, S., & Spanjers, D. (2010). *Characteristics of reading aloud, word identification, and maze selection as growth measures: Relationship between growth and criterion measures (RIPM Technical Report 20)*. Retrieved from University of Minnesota, Research Institute on Progress Monitoring website: http://www.progressmonitoring.net

*Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice, 25*(2), 60-75.

*Farmer, E. (2013). *Examining predictive validity and rates of growth in curriculum-based measurement with English language learners in the intermediate grades* (Unpublished doctoral dissertation). Loyola University Chicago, Chicago: IL.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). Learning disabilities: From identification to intervention. New York: Guilford.

*Ford, L. A. (2008). *The Relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency and the Terra Nova, 2nd Ed. performance on Ohio grade 3 reading achievement assessment* (Unpublished master's thesis). Marshall University, Huntington: WV.

*Fore, C., Boon, R. T., & Martin, C. (2007). Concurrent and predictive criterion-related validity of curriculum-based measurement for students with emotional and behavioral disorders. *International Journal of Special Education, 22*(2), 24-32.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45-58.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*, 7-21.

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

*Galloway, T. W. (2010). *Oral reading fluency and maze measures as predictors of performance on North Carolina End-of-Grade Assessment of Reading Comprehension*. (Unpublished doctoral dissertation). The University of North Carolina at Charlotte.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7,* 6-10.

*Graney, S. B., Martinez, R. S., Missall, K. N., Aricak, T. (2010). Universal screening of reading in late elementary school: R-CBM versus CBM maze. *Remedial and Special Education, 31*, 368-377.

Goodwin, L. D., & Reach, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *The Journal of Experimental Education, 74*, 251-266.

Hamilton, C. R., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgements of reading comprehension and oral reading skills. *School Psychology Review, 32,* 228-240.

Harcourt Educational Measurement. (2003). *Stanford Achievement Test* (10th ed.). San Antonio, TX: Author.

Harwell, M., & Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *The journal of experimental education*, *76*(4), 403-430.

Harwell, M., Maeda, Y., Bishop, K., & Xie, A. (2016). The Surprisingly modest relationship between SES and educational achievement. *The Journal of Experimental Education*, 1-18.

Hedges L. V., & Vevea J. L. (1998). Fixed- and Random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486-504.

*Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the classification accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372-386.

*Hixson, M. D., & McGlinchey, M. T. (2004). The relationship between race, income, and oral reading fluency and performance on two reading comprehension measures. *Journal of Psychoeducational Assessment, 22,* 351-364.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A. (2005). Iowa Test of Basic Skills, Form M. Itasca, IL: Riverside Publishing.

*Hunley, S. A., Davies, S. C., & Miller, C. R. (2013). The relationship between curriculum-based measures in oral reading fluency and high-stakes tests for seventh grade students. *RMLE Online*, *36*(5), 1-8.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*, 421-432.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for service delivery in an RTI framework: Candidate measures. *School Psychology Review, 36*, 582-599.

*Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*(3), 374.

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

*Kim, Y. S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, *102*(3), 652.

*Kim, Y. S., Petscher, Y., & Foorman, B. (2015). The unique relation of silent reading fluency to end-of-year reading comprehension: Understanding individual differences at the student, classroom, school, and district levels. *Reading and Writing, 28*(1), 131-150.

*Kloo, A. M. (2007). *The decision-making utility and predictive power of DIBELS for students' reading achievement in Pennsylvania's Reading first schools* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh: PA.

*Kranzler, J. H., & Miller, D. M., & Jordan, L. (1999). An examination of Racial/Ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly, 14*, 327-342.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293-323.

*LeRoux, M. J. (2010). *Using curriculum-based measurement to predict eighth-grade student performance on a statewide reading assessment*. (Unpublished doctoral dissertation). University of Oregon.

Lesauxr, N. K., & Kieffer, M. J. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal, 47*, 596-632.

Maeda, Y., & Harwell, M. R. (2016). Guidelines for using the Q test in meta-analysis. *Mid-Western Educational Researcher, 28*(1), 55-72.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47,* 315-335.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-Based Measurement: Assessing Special Children,* (pp. 18-78). New York: Guilford.

*McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193-203.

McMaster, K., Wayman, 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*, 17-31.

*Megert, B. R. (2010). *Establishing predictive validity for oral passage reading fluency and vocabulary curriculum-based measures (CBMs) for sixth grade students*. (Unpublished doctoral dissertation). University of Oregon, Eugene: OR.

*Munger, K. A., & Blachman, B. A. (2013). Taking a "simple view" of the dynamic indicators of basic early literacy skills as a predictor of multiple measures of third-grade reading comprehension. *Psychology in the Schools*, *50*(7), 722-737.

Nese, J. F., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment. *The Elementary School Journal, 111*, 608-624.

O'Connor. R. E., & Jenkins. J. R. (1999). The prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3*, 159-197.

*Paleologos, T. M., & Brabham, E. G. (2011). The effectiveness of DIBELS oral reading fluency for predicting reading comprehension of high- and low-income students. *Reading Psychology, 32*, 54-74.

Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26*, 195-218.

*Pearce, L. R., & Gayle, R. (2008). Oral reading fluency as a predictor of reading comprehension on a state's measure of adequate yearly progress. *International Journal of Psychology: A Biopsychosocial Approach, 1*, 51-70.

*Pearce, L. R., & Gayle, R. (2009). Oral reading fluency as a predictor of reading comprehension with American Indian and White elementary students. *School Psychology Review, 38*, 419-427.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. *The science of reading: A handbook* (pp. 227-247). Oxford, UK: Blackwell.

*Petscher, Y., Kim, Y., & Foorman, B. R. (2011). The Importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention, 36,* 158-166.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427-469.

*Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal*, *48*(2), 462-501.

*Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*, 546-567.

*Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*, 343-366.

Rosenthal, D. A., Hoyt, W. T., Ferrin, J. M., Miller, S., & Cohen, N. D. (2006). Advanced methods in meta-analytic research: Applications and implications for rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 49*(4), 234–246.

*Sáez, L., Park, B., Nese, J. F., Jamgochian, E., Lai, C. F., Anderson, D., ... & Tindal, G. (2010). *Technical Adequacy of the easyCBM Reading Measures (Grades 3-7), 2009-2010 Version* (Technical Report #1005). Eugene, OR: Behavioral Research and Teaching.

*Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal, 107*, 429-448.

*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.

*Shapiro, E. S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and individual differences*, *18*(3), 316-328.

*Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-Based Measures and performance on state assessment and standardized tests: Reading and: math problem in Pennsylvania, *Journal of Psychoeducational Assessment, 24*, 19-36.

*Shaw, R., & Shaw, D. (2002). *DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP*) (Technical report). Eugene, OR: University of Oregon.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34*, 164-172.

Shinn, M. R. (1989). Curriculum-based measurement: Assessing special children. New York: Guilford Press.

*Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43*, 527-535.

*Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23,* 304-325.

Snow, C. (2002). RAND Reading Study Group. Reading for understanding. Santa Monica, CA: RAND.

*Spear-Swerling, Louise. (2006). Children's reading comprehension and oral reading fluency in easy text. *Reading and Writing*, *19*, 199-220.

*Stage, S. A., & Jacobsen,M. D. (2001). Predicting student success on a state mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*, 407-419.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*, 795-819.

*Strokes, N. O. (2010). *Examining the relationship among reading curriculum-based measures, level of language proficiency, and state accountability test scores with middle school Spanish-speaking English language learners*. (Unpublished doctoral dissertation). Loyola University Chicago, Chicago: IL.

*Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures. *Learning Disabilities Research & Practice, 24*(3), 132-142.

Thurlow, R., & van den Broek, P. (1997). Automaticity and inference generation during reading comprehension. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*(2), 165-181.

*Uribe-Zarain, X. (2006). *Relationship between student performance on DIBELS oral reading fluency and third grade reading DSTP* (Technical report). Newark, DE: University of Delaware Education Research & Development Center.

*Utchell, L. A. (2011). *Relationships among Early Literacy Curriculum Based Measurement and Reading State Criterion Tests over Time* (Unpublished doctoral dissertation). Duquesne University, Pittsburgh: PA.

*Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, *45*(3), 270-291.

van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes: A Multidisciplinary Journal, 39*, 299-316.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw*, *36*(3), 1-48.

*Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A. L., & Murray, C. S. (2010). Differences in the relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention*, *35*, 67-77.

Wayman, M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education*, *41*, 85-120.

Wiederholt, J. L., & Bryant, B. R. (2001). Gray oral reading test-(GORT-4). *Austin, TX: Pro-Ed*.

*Wiley, H., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*, 207-214.

*Wilson, J. (2005). *The relationship of dynamic indicators of basic early literacy skills (DIBELS) oral reading fluency to performance on Arizona instrument to measure standards (AIMS)* (Technical report). Tempe, AZ: Tempe School District.

*Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11*, 85-104.

Yeo, S. (2008). *Relation between 1-minute CBM reading aloud measure and reading comprehension tests: A multilevel meta-analysis*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*, 412-422.

Appendix A

Summary of Included Studies

Table A1

Summary of Included Studies (Primary Grades)

Study	Publication type	Participants		CBM				State test			Time interval (month)	Corr
		Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format		
Cook (2003)	dissertation	1	79	OR	standard	3	teacher	SAT-9	commercial	multiple	0	.73
Devena (2014)	dissertation	1	312	OR	standard	3	teacher	AIMS	state	multiple	24	0.46
Kloo (2006)	dissertation	1	8,595	OR	standard	3	na	PSSA	state	mixed	28	0.40
Munger &Blachman (2013)	journal	1	35	OR	standard	3	teacher	NYSELA 3	state	mixed	19	0.56
Reidel & Samuels (2007)	journal	1	1,518	OR	standard	3	teacher	Terra Nova	commercial	multiple	14	0.54
Schilling et al. (2007)	journal	1	2,588	OR	standard	3	teacher	ITBS	commercial	multiple	0	0.74
Wanzek et al. (2010)	journal	1	270	OR	standard	3	researcher	SAT-10	commercial	multiple	0	0.64
Devena (2014)	dissertation	2	312	OR	standard	3	teacher	AIMS	state	multiple	12	0.45
Echols (2010)	dissertation	2	982	OR	standard	3	teacher	WASL	state	mixed	12	0.65
Keller-Margulis et al. (2008)	journal	2	150	OR	standard	1	teacher	PSSA	state	mixed	13	0.71
Kranzler et al. (1999)	journal	2	84	OR	RD	6	researcher	CAT	commercial	multiple	1	0.63
Petscher et al. (2011)	journal	2	17,778	OR	standard	1	na	SAT-10	commercial	multiple	8	0.64
Schilling et al. (2007)	journal	2	2,437	OR	standard	3	teacher	ITBS	commercial	multiple	0	0.75
Valencia et al. (2010)	journal	2	93	OR	RD	2	researcher	ITBS	commercial	multiple	9	0.55

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A1 (cont.)

Summary of Included Studies (Primary Grades)

		Participants		CBM				State test				
Study	Publication type	Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format	Time interval (month)	Corr
Wanzek et al. (2010)	journal	2	347	OR	standard	3	researcher	SAT-10	commercial	multiple	0	0.68
Acquavita (2012)	dissertation	3	1,663	OR	standard	3	teacher	FCAT-SSS	state	multiple	0	0.61
Anderson, Alonzo, Tindal (2011)	tech report	3	3888	OR	standard	1	teacher	OAKS	state	mixed	0	0.67
Ardoin et al. (2004)	journal	3	77	OR	standard	1	researcher	ITBS	commercial	multiple	2.33	0.58
Ardoin et al. (2004)	journal	3	75	Maze	standard	1	researcher	ITBS	commercial	multiple	2.33	0.49
Baker et al. (2008)	journal	3	2,400	OR	standard	3	na	OAKS	state	mixed	0	0.67
Barger (2003)	tech report	3	38	OR	standard	3	na	North Carolina	state	multiple	0.25	0.73
Buck & Torgesen (2003)	tech report	3	1102	OR	standard	3	teacher	FCAT-SSS	state	multiple	0	0.70
Canto (2006)	dissertation	3	186	OR	standard	3	teacher	FCAT-SSS	state	multiple	1	0.68
Crawford et al. (2001)	journal	3	51	OR	standard	3	teacher	OAKS	state	mixed	2	0.60
Devena (2014)	dissertation	3	312	OR	standard	3	teacher	AIMS	state	multiple	0	0.67
Espin et al. (2010)	tech report	3	41	OR	standard	1	researcher	MCA	state	mixed	3	0.54
Espin et al. (2010)	tech report	3	41	Maze	standard	1	teacher	MCA	state	mixed	3	0.49
Ford (2008)	dissertation	3	136	OR	standard	3	researcher	Ohio reading	state	mixed	1	0.62

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A1 (cont.)

Summary of Included Studies (Primary Grades)

Study	Publication type	Participants		CBM				State test			Time interval (month)	Corr
		Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format		
Galloway (2010)	dissertation	3	110	OR	standard	3	researcher	North Carolina reading	state	multiple	1	0.56
Galloway (2010)	dissertation	3	110	Maze	standard	3	teacher	North Carolina reading	state	multiple	1	0.53
Hintze & Silberglitt (2005)	journal	3	1,766	OR	RD	3	teacher	MCA	state	mixed	0	0.69
Kim et al. (2010)	journal	3	12,536	OR	standard	3	na	SAT-10	commercial	multiple	2	0.10
Kim et al. (2015)	journal	3	156,179	Maze	standard	2	teacher	FCAT	state	multiple	6	0.66
Kloo (2006)	dissertation	3	8,317	OR	standard	3	na	PSSA	state	mixed	0	0.71
Kranzler et al. (1999)	journal	3	76	OR	RD	6	researcher	CAT	commercial	multiple	1	0.52
Paleologos & Brabham (2011)	journal	3	56	OR	standard	1	teacher	SAT-10	commercial	multiple	0	0.23
Pearce & Gayle (2008)	journal	3	544	OR	standard	1	na	Dstep	state	multiple	3	0.63
Pearce & Gayle (2009)	journal	3	543	OR	standard	1	teacher	Dstep	state	multiple	3	0.63
Roehrig et al. (2008)	journal	3	16539	OR	standard	1	na	FCAT-SSS	state	multiple	0	0.71
Saez et al. (2010)	tech report	3	2216	OR	standard	1	teacher	OAKS	state	mixed	0	0.67
Schilling et al. (2007)	journal	3	2,527	OR	standard	3	teacher	ITBS	commercial	multiple	0	0.63

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A1 (cont.)

Summary of Included Studies (Primary Grades)

		Participants		CBM				State test				
Study	Publication type	Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format	Time interval (month)	Corr
Shapiro et al. (2006)	journal	3	185	OR	standard	1	teacher	PSSA	state	mixed	0	0.67
Shapiro et al. (2008)	journal	3	401	OR	standard	3	teacher	PSSA	state	mixed	2	0.78
Shapiro et al. (2014)	journal	3	158	OR	standard	3	researcher	PSSA	state	mixed	0	0.57
Shaw & Shaw (2002)	tech report	3	52	OR	standard	3	teacher	CSAP	state	mixed	0	0.80
Silberglitt & Hintze (2005)	journal	3	2,191	OR	standard	3	na	MCA	state	mixed	0	0.71
Silberglitt et al. (2006)	journal	3	3,165	OR	standard	3	teacher	MCA	state	mixed	2	0.68
Sledge-Murphy (2011)	dissertation	3	1,094	OR	standard	3	teacher	iLEAP	state	mixed	7	0.56
Spear-Swerling (2006)	journal	3	61	OR	standard	2	na	Conneticut Mastery Test	state	mixed	10	0.65
Uribe-Zarain (2006)	dissertation	3	630	OR	standard	3	na	DSTP	state	mixed	2	0.61
Utchell (2011)	dissertation	3	130	OR	standard	3	teacher	PSSA	state	mixed	27	0.46
Wanzek et al. (2010)	journal	3	461	OR	standard	3	researcher	SAT-10	commercial	multiple	0	0.69
Wiley & Deno (2005) – EL	journal	3	15	OR	standard	3	teacher	MCA	state	mixed	6	0.61
Wiley & Deno (2005) – non-EL	journal	3	21	OR	standard	3	teacher	MCA	state	mixed	6	0.71
Wiley & Deno (2005) – EL	journal	3	15	Maze	standard	3	teacher	MCA	state	mixed	6	0.52
Wiley & Deno (2005) – non-EL	journal	3	21	Maze	standard	3	teacher	MCA	state	mixed	6	0.73

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A1 (cont.)

Summary of Included Studies (Primary Grades)

Study	Publication type	Participants		CBM				State test			Time interval (month)	Corr
		Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format		
Wilson (2005)	tech report	3	240	OR	standard	3	na	AIMS	state	multiple	0	0.74
Wood (2006) 1	journal	3	82	OR	standard	3	teacher	CSAP	state	mixed	2	0.70
Paleologos & Brabham (2011) 1	journal	3	56	OR	standard	1	teacher	SAT-10	commercial	multiple	0	0.60

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A2

Summary of Included Studies (Intermediate Grades)

Study	Publication type	Participants		CBM				State test			Time interval (month)	Corr
		Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format		
Anderson, Alonzo, Tindal (2011)	tech report	4	3740	OR	standard	1	teacher	OAKS	state	mixed	0	0.66
Farmer (2013)	dissertation	4	362	OR	standard	1	teacher	ISAT	state	mixed	2	0.72
Galloway (2010)	dissertation	4	111	OR	standard	3	researcher	North Carolina reading	state	multiple	1	0.66
Galloway (2010)	dissertation	4	111	Maze	standard	3	teacher	North Carolina reading	state	multiple	1	0.62
Graney et al. (2010)	journal	4	76	OR	standard	3	researcher	ISTEP	state	mixed	12	0.72
Graney et al. (2010)	journal	4	76	Maze	standard	1	researcher	ISTEP	state	mixed	12	0.67
Hixson & McGlinchey (2004)	journal	4	442	OR	RD	1	teacher	MEAP	state	multiple	0.5	0.54
Keller-Margolis et al. (2008)	journal	4	150	OR	standard	1	teacher	PSSA	state	mixed	13	0.69
Kim et al. (2015)	journal	4	140,045	Maze	standard	2	teacher	FCAT	state	multiple	6	0.66
Kranzler et al. (1999)	journal	4	94	OR	RD	6	researcher	CAT	commercial	multiple	1	0.54
McGlinchey & Hixson (2004)	journal	4	1,362	OR	RD	1	teacher	MEAP	state	multiple	0.5	0.67
Pearce & Gayle (2008)	journal	4	265	OR	standard	1	na	Dstep	state	multiple	3	0.66
Reis et al. (2011)	journal	4	1,192	OR	RD	3	researcher	ITBS	commercial	multiple	5	0.80
Saez et al. (2010)	tech report	4	2265	OR	standard	1	teacher	OAKS	state	mixed	0	0.66
Shapiro et al. (2008)	journal	4	394	OR	standard	3	teacher	PSSA	state	mixed	2	0.68

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A2 (cont.)

Summary of Included Studies (Intermediate Grades)

Study	Publication type	Participants		CBM				State test			Time interval (month)	Corr
		Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format		
Stage & Jacobson (2001)	journal	4	173	OR	RD	1	teacher	WASL	state	mixed	0	0.44
Valencia et al. (2010)	journal	4	91	OR	RD	2	researcher	ITBS	commercial	multiple	9	0.48
Wood (2006)	journal	4	101	OR	standard	3	teacher	CSAP	state	mixed	2	0.67
Anderson, Alonzo, Tindal (2011)	tech report	5	3851	OR	standard	1	teacher	OAKS	state	mixed	0	0.65
Espin, Deno et al. (2010)	tech report	5	32	OR	standard	1	researcher	MCA	state	mixed	3	0.39
Espin, Deno et al. (2010)	tech report	5	32	Maze	standard	1	teacher	MCA	state	mixed	3	0.30
Farmer (2013)	dissertation	5	339	OR	standard	1	teacher	ISAT	state	mixed	2	0.65
Galloway (2010)	dissertation	5	115	OR	standard	3	researcher	North Carolina reading	state	multiple	1	0.75
Galloway (2010)	dissertation	5	115	Maze	standard	3	teacher	North Carolina reading	state	multiple	1	0.52
Kim et al. (2015)	journal	5	140,533	Maze	standard	2	teacher	FCAT	state	multiple	6	0.65
Kranzler et al. (1999)	journal	5	72	OR	RD	6	researcher	CAT	commercial	multiple	1	0.51
Pearce & Gayle (2008)	journal	5	235	OR	standard	1	na	Dstep	state	multiple	3	0.66
Saez et al. (2010)	tech report	5	2265	OR	standard	1	teacher	OAKS	state	mixed	0	0.65

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A2 (cont.)

Summary of Included Studies (Intermediate Grades)

Study	Publication type	Participants		CBM				State test			Time interval (month)	Corr
		Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format		
Shapiro et al. (2006)	journal	5	206	OR	standard	1	teacher	PSSA	state	mixed	0	0.66
Shapiro et al. (2008)	journal	5	205	OR	standard	3	teacher	PSSA	state	mixed	2	0.75
Shapiro et al. (2014)	journal	5	113	OR	standard	3	researcher	PSSA	state	mixed	0	0.46
Silberglitt et al. (2006)	journal	5	3,283	OR	standard	3	teacher	MCA	state	mixed	2	0.65
Wiley & Deno (2005)	journal	5	14	OR	standard	3	teacher	MCA	state	mixed	6	0.69
Wiley & Deno (2005)	journal	5	19	OR	standard	3	teacher	MCA	state	mixed	6	0.57
Wiley & Deno (2005)	journal	5	14	Maze	standard	3	teacher	MCA	state	mixed	6	0.57
Wiley & Deno (2005)	journal	5	19	Maze	standard	3	teacher	MCA	state	mixed	6	0.73
Wood (2006)	journal	5	98	OR	standard	1	teacher	CSAP	state	mixed	2	0.75
Alonzo & Tindal (2004)	tech report	6	263	OR	standard	1	teacher	OSA in reading	state	mixed	na	0.54
Anderson, Alonzo, Tindal (2011)	tech report	6	3862	OR	standard	1	teacher	OAKS	state	mixed	0	0.67
Farmer (2013)	dissertation	6	355	OR	standard	1	teacher	ISAT	state	mixed	2	0.73
Kim et al. (2015)	journal	6	128,556	Maze	standard	2	teacher	FCAT	state	multiple	6	0.64
Megert (2010)	dissertation	6	678	OR	standard	1	teacher	OAKS	state	mixed	7	0.64
Saez et al. (2010)	tech report	6	1190	OR	standard	1	teacher	OAKS	state	mixed	0	0.67
Strokes (2010)	dissertation	6	349	OR	standard	3	na	AIMS	state	multiple	7	0.58
Strokes (2010)	dissertation	6	349	Maze	standard	1	na	AIMS	state	multiple	0	0.41
Valencia et al. (2010)	journal	6	95	OR	RD	2	researcher	ITBS	commercial	multiple	9	0.48

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A3

Summary of Included Studies (Secondary Grades)

Study	Publication type	Participants		CBM				State test			Time interval (month)	Corr
		Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format		
Anderson, Alonzo, Tindal (2011)	tech report	7	3600	OR	standard	1	teacher	OAKS	state	mixed	0	0.32
Baker et al. (2015)	journal	7	1,481	OR	standard	3	teacher	OAKS Reading	state	mixed	2	0.69
Decker et al. (2014)	journal	7	83	OR	RD	3	teacher	MEAP	state	multiple	1	0.51
Decker et al. (2014)	journal	7	83	Maze	RD	1	teacher	MEAP	state	multiple	1	0.54
Denton et al. (2011)	journal	7	1,421	OR	standard	5	researcher	TAKS	state	multiple	7	0.50
Denton et al. (2011)	journal	7	1,421	Maze	standard	1	researcher	TAKS	state	multiple	7	0.40
Fore et al. (2007)	journal	7	50	OR	standard	1	na	CRCT	state	multiple	na	0.40
Fore et al. (2007)	journal	7	50	Maze	standard	1	na	CRCT	state	multiple	na	0.44
Hunley et al. (2013)	journal	7	75	OR	RD	3	researcher	Ohio reading test	state	mixed	1	0.76
Kim et al. (2015)	journal	7	127,030	Maze	standard	2	teacher	FCAT	state	multiple	6	0.65
Saez et al. (2010)	tech report	7	2428	OR	standard	1	teacher	OAKS	state	mixed	0	0.69
Silberglitt et al. (2006)	journal	7	528	OR	standard	3	teacher	MCA	state	mixed	2	0.60
Silberglitt et al. (2006)	journal	7	528	Maze	standard	1	teacher	MCA	state	mixed	2	0.54
Baker et al. (2015)	journal	8	1,462	OR	standard	3	teacher	OAKS Reading	state	mixed	2	0.69
Decker et al. (2014)	journal	8	95	OR	RD	3	teacher	MEAP	state	multiple	1	0.63
Decker et al. (2014)	journal	8	95	Maze	RD	1	teacher	MEAP	state	multiple	1	0.58

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Table A3 (cont.)

Summary of Included Studies (Secondary Grades)

		Participants		CBM				State test				
Study	Publication type	Gr	*n*	CBM Type	Develop type	# of passage	admin	Test name	Develop type	Format	Time interval (month)	Corr
Espin et al. (2010)	journal	8	236	OR	RD	1	researcher	MBST	state	multiple	5	0.78
Espin et al. (2010)	journal	8	236	Maze	RD	1	researcher	MBST	state	multiple	5	0.78
Kim et al. (2015)	journal	8	129,341	Maze	standard	2	teacher	FCAT	state	multiple	6	0.60
LeRoux (2010)	dissertation	8	126	OR	standard	3	teacher	OAKS	state	mixed	1	0.64
LeRoux (2010)	dissertation	8	160	OR	standard	3	teacher	OAKS	state	mixed	1	0.58
LeRoux (2010)	dissertation	8	126	Maze	standard	1	teacher	OAKS	state	mixed	1	0.58
LeRoux (2010)	dissertation	8	160	Maze	standard	1	teacher	OAKS	state	mixed	1	0.54
Ticha et al. (2009)	journal	8	35	OR	RD	1	researcher	MBST	state	multiple	0	0.77
Ticha et al. (2009)	journal	8	35	Maze	RD	1	researcher	MBST	state	multiple	0	0.82
Kim et al. (2015)	journal	9	131,629	Maze	standard	2	teacher	FCAT	state	multiple	6	0.57
Kim et al. (2015)	journal	10	124,864	Maze	standard	2	teacher	FCAT	state	multiple	6	0.57

Note. Gr = Grade; OR = oral reading; standard = standardized; RD = researcher-developed; admin = administrator; Corr = correlation.

Appendix B

Student Demographic Information by Study

Table B1

Student Demographic Information

Study[a]	Characteristic of sample (proportion)				
	Female	ELL	Special education	FRL	White
Acquavita (2012)	0.48	na	na	0.86	0.08
Alonzo & Tindal (2004)	0.50	0.02	0.07	na	0.68
Anderson, Alonzo, Tindal (2011)	0.48	0.07	0.16	0.51	0.62
Ardoin et al. (2004)	0.55	na	na	0.44	0.58
Baker et al. (2008)	na	0.32	0.10	0.69	na
Baker et al. (2015)	0.49	0.07	0.11	0.49	0.56
Barger (2003)	na	na	na	na	na
Buck & Torgesen (2003)	0.49	0.01	0.19	0.46	0.83
Canto (2006)	0.60	na	na	0.64	0.35
Cook (2003)	0.51	na	na	0.57	1
Crawford et al. (2001)	0.57	na	0.18	na	0.94
Decker et al. (2014)	na	na	na	0.19	0.95
Denton et al. (2011)	0.52	0	0	0.63	0.19
Devena (2014)	0.49	0.03	0.14	na	0.56
Echols (2010)	0.46	0.22	0.12	na	0.30
Espin et al. (2010)	0.57	0.42	0.09	0.58	0.34
Espin, Deno et al. (2010)	na	0.08	0.07	na	0.60
Farmer (2013)	na	0.21	0.13	0.52	0.34
Ford (2008)	0.47	na	0.18	0.63	1
Fore et al. (2007)	0.26	na	1	0	0.40
Galloway (2010)	na	na	0.14	0.34	0.79
Graney et al. (2010)	na	na	na	0.24	0.93
Hintze & Silberglitt (2005)	0.49	na	0.05	0.30	0.94
Hixson & McGlinchey (2004)	na	na	na	0.52	0.55
Hunley et al. (2013)	na	na	0.17	0.20	na
Keller-Margulis et al. (2008)	na	0.08	na	0.32	0.58
Kim et al. (2010)	0.50	na	na	0.31	0.41
Kim et al. (2015)	0.49	0.07	0.15	0.60	0.45
Kloo (2006)					

Note. [a] = studies were arranged in alphabetical order; ELL = English language learner; FRL = free or reduced lunch; na = not available.

Table B1 (cont.)

Characteristics of Study Samples

Study[a]	Characteristic of sample (proportion)				
	Female	ELL	Special education	FRL	White
Kranzler et al. (1999)	0.48	0	0	na	0.73
LeRoux (2010)	0.50	na	0.16	0.52	0.69
McGlinchey & Hixson (2004)	0.48	na	0.06	0.64	0.49
Megert (2010)	0.48	na	0.16	na	0.78
Munger & Blachman (2013)	na	0.17	0.25	0.73	0.26
Paleologos & Brabham (2011)	na	na	na	0.50	na
Pearce & Gayle (2008)	0.47	na	na	na	0.77
Pearce & Gayle (2009)	0.48	0.02	0.14	0.41	0.79
Petscher, Kim, & Foorman (2011)	0.50	0.15	0.11	0.77	0.4
Reidel & Samuels (2007)	0.50	0.04	0	0.85	0.07
Reis et al. (2011)	na	0.12	0.13	0.59	0.33
Roehrig et al. (2008)	0.49	0.12	0.17	0.75	0.36
Saez et al. (2010)	0.48	0.04	0.19	0.41	0.68
Schilling et al. (2007)	na	0.16	0.09	0.81	0.25
Shapiro et al. (2006)	na	0.08	0.11	0.32	na
Shapiro et al. (2008)	na	na	na	0.39	0.65
Shapiro et al. (2014)	na	na	0	0.31	0.79
Shaw & Shaw (2002)	na	na	na	na	na
Silberglitt & Hintze (2005)	0.47	na	na	0.26	0.95
Silberglitt et al. (2006)	0.49	na	na	0.11	0.94
Sledge-Murphy (2011)	na	na	na	0.55	na
Spear-Swerling (2006)	0.52	0	0	na	na
Stage & Jacobson (2001)	0.46	na	0.06	0.15	0.90
Strokes (2010)	na	0.26	na	na	0.07
Ticha et al. (2009)	0.57	0.03	0.14	0.40	0.49
Uribe-Zarain (2006)	0.50	0.03	0.15	0.59	0.44
Utchell (2011)	na	0.01	0.15	0.25	0.95
Valencia et al. (2010)	na	0.33	na	0.43	0.46
Wanzek et al. (2010)	0.50	0	0.19	0.74	0.13
Wiley & Deno (2005)	0.49	0.50	na	0.82	na
Wilson (2005)	0.45	0.27	na	0.70	0.34
Wood (2006)	na	0	0.09	na	0.86

Note. [a] = studies were arranged in alphabetical order; ELL = English language learner; FRL = free or reduced lunch; na = not available.

Appendix C

Summary of State Tests Used

Table C1

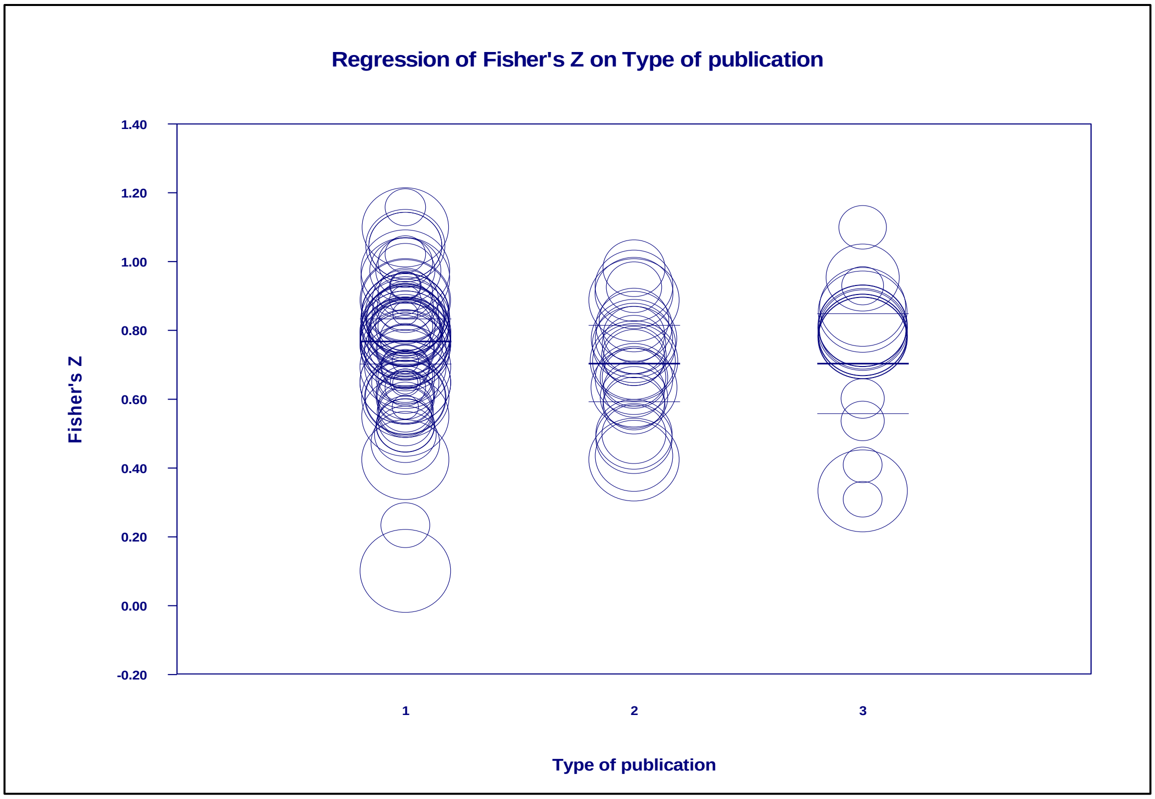
Summary of State Tests Used

Name of state test	State	Number of studies
1. AIMS (Arizona's Instrument to Measure Standards)	Arizona	3
2. CAT (California Achievement Test)	California	1
3. CMT (Connecticut Mastery Test)	Connecticut	1
4. CRCT (Criterion-Referenced Competency Tests)	Georgia	1
5. CSAP (Colorado Student Assessment Program)	Colorado	3
6. Dstep (Dakota State Test of Educational Progress)	South Dakota	2
7. DSTP (Delaware Student Testing Program)	Delaware	1
8. FCAT (Florida Comprehensive Assessment Test)	Florida	5
9. iLEAP (LEAP Alternate Assessment)	Louisiana	1
10. ISAT (Illinois Standards Achievement Test)	Illinois	1
11. ISTEP (Indiana Statewide Testing for Educational Progress)	Indiana	1
12. ITBS (Iowa Test of Basic Skills)	Iowa	4
13.MBST (Minnesota Basic Skills Test)	Minnesota	2
14. MCA (Minnesota Comprehensive Assessments Series)	Minnesota	5
15. MEAP (Michigan Educational Assessment Program)	Michigan	3
16. North Carolina End of Grade Reading	North Carolina	2
17. NYSELA (New York State English and Language Arts Test)	New York	1
18. OAKS (Oregon Statewide Assessment System)	Oregon	6
19. OSRA (Oregon Statewide Reading Assessment)	Oregon	1
20. Ohio Achievement Assessments	Ohio	1
21. PSSA (Pennsylvania System of School Assessment)	Pennsylvania	6
22. SAT-9/10 (Stanford Achievement Test)	-	5
23. TAKS (Texas Assessment of Knowledge and Skills)	Texas	1
24. TerraNova	Alaska	1
25. WASL (Washington Assessment of Student Learning)	Washington	2

Appendix D

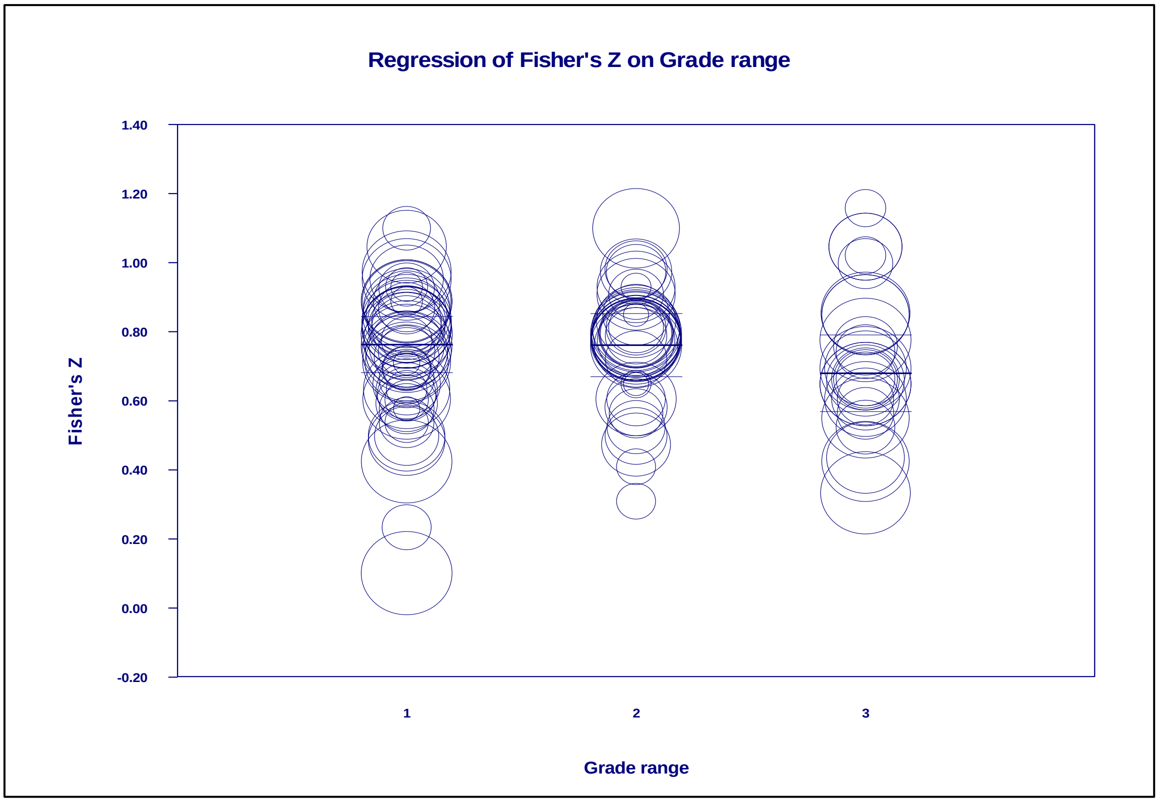
Scatterplots of Meta-regression: By Moderators

Figure D1. Scatter Plot of Meta-regression on Type of Publication



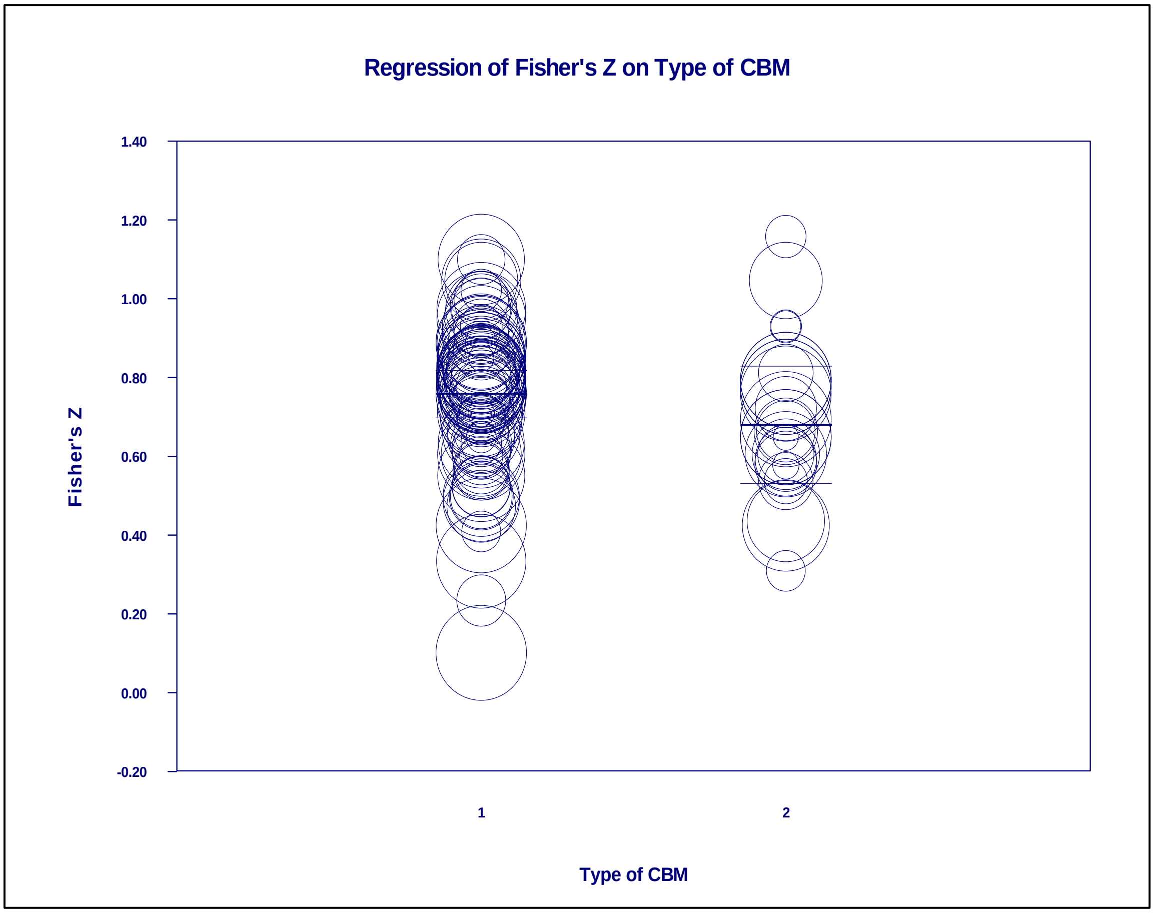
Note. 1 = journal articles; 2 = dissertations; 3 = technical reports.

Figure D2. Scatter Plot of Meta-regression on Grade Range



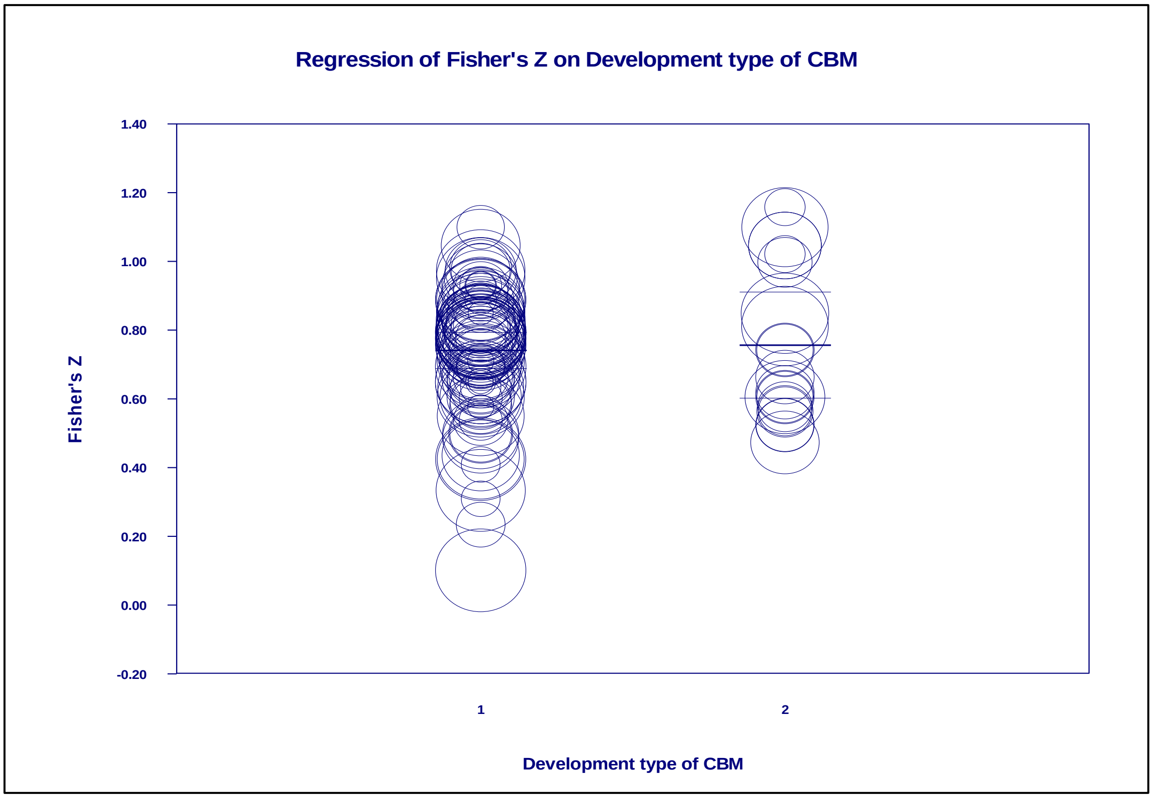
Note. 1 = primary grades; 2 = intermediate grades; 3 = secondary grades.

Figure D3. Scatter Plot of Meta-regression on Type of CBM



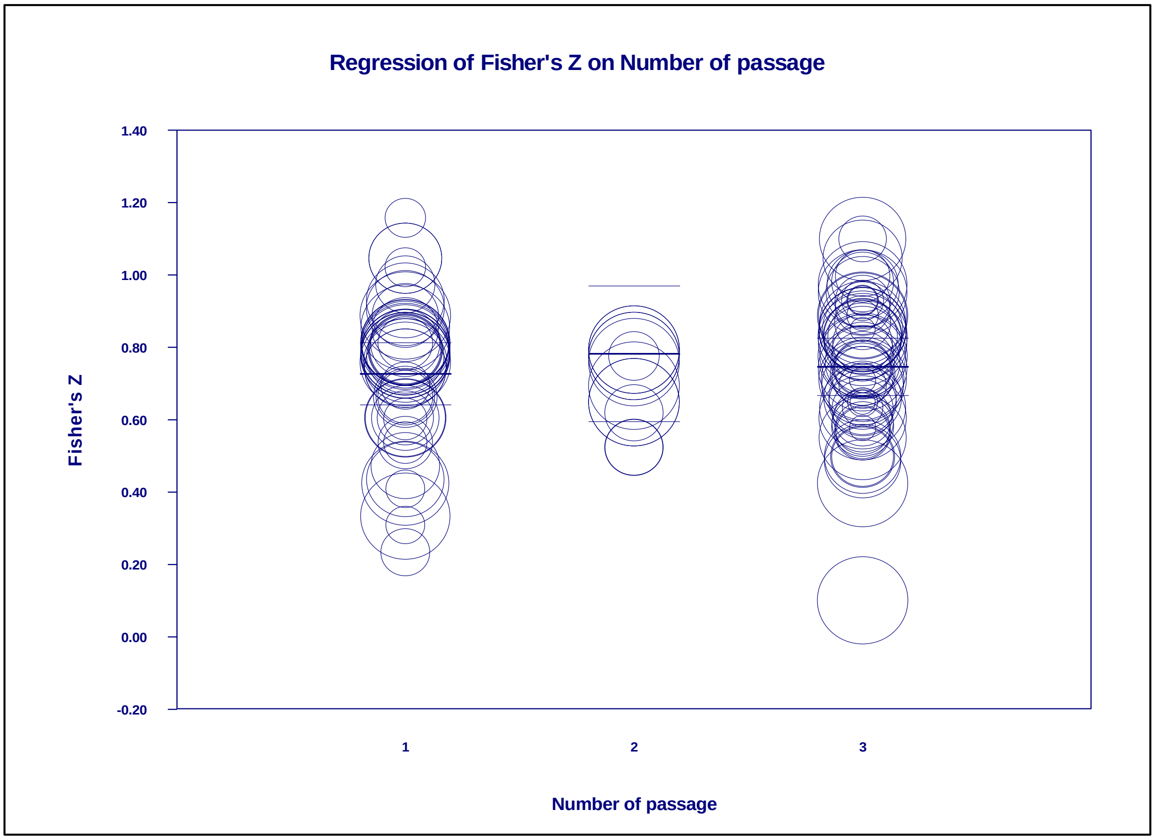
Note. 1 = oral reading; 2 = maze.

Figure D4. Scatter Plot of Meta-regression on Development Type of CBM



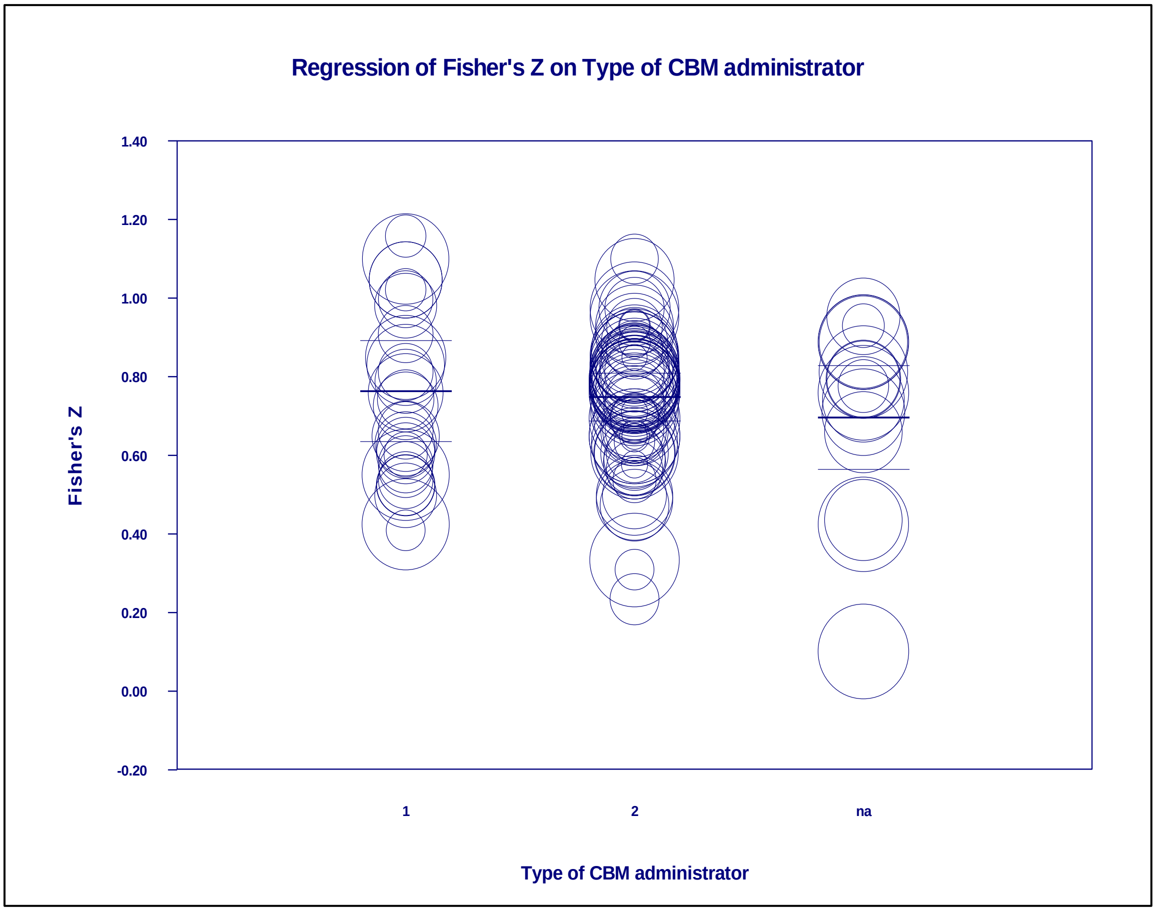
Note. 1 = standardized; 2 = researcher-developed.

Figure D5. Scatter Plot of Meta-regression on Number of CBM Passage



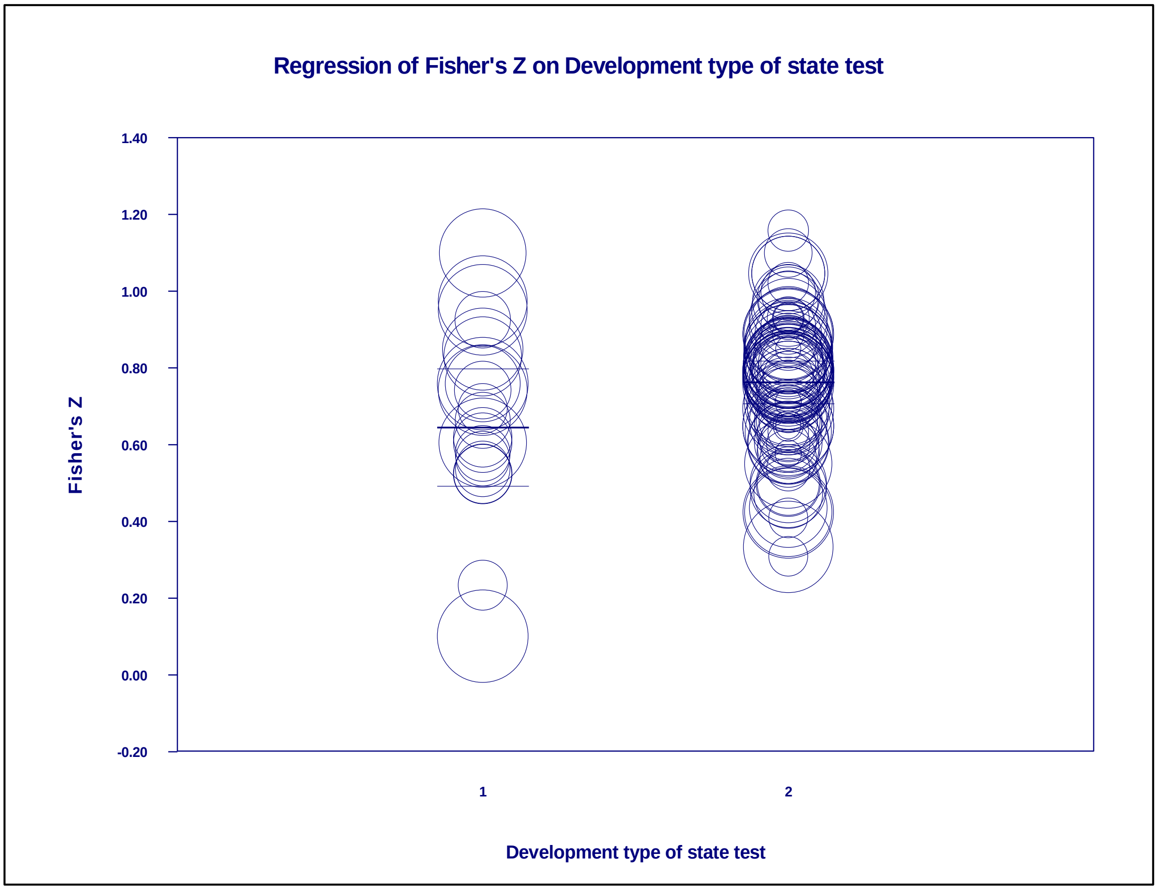
Note. 1 = one passage; 2 = two passages; 3 = three passages.

Figure D6. Scatter Plot of Meta-regression on Type of CBM Administrator



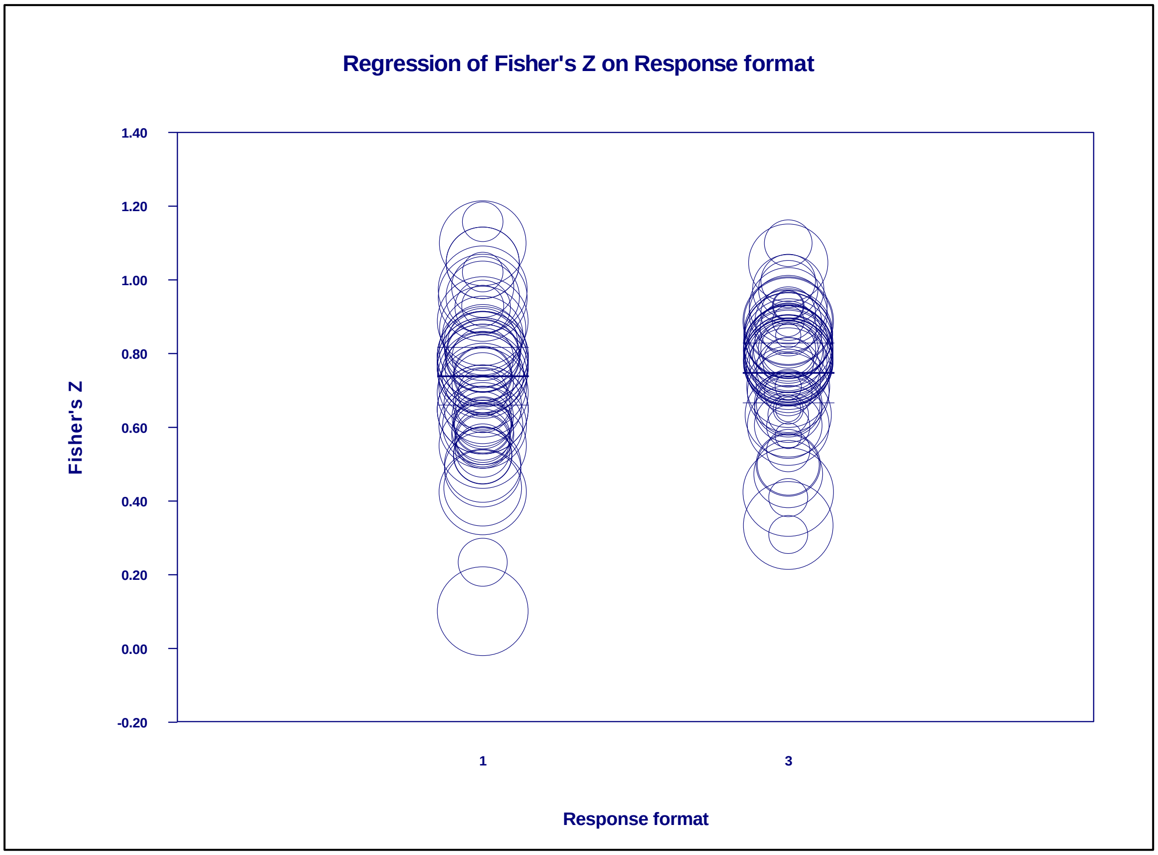
Note. 1 = researcher (including graduate assistant); 2 = school personnel (teacher, school psychologist); NA = not available.

Figure D7. Scatter Plot of Meta-regression on Development Type of State Test



Note. 1 = commercially developed; 2 = state developed.

Figure D8. Scatter Plot of Meta-regression on Response Format



Note. 1 = multiple choice; 2 = mixed (multiple choice and open ended).